

大数据·数据管理与数据工程

◎ 赵眸光 赵勇 编著



清華大学出版社

大数据·数据管理与数据工程

赵眸光　赵　勇　编著

清华大学出版社
北　京

内 容 简 介

大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的必然产物，必将对物联网产业产生深远的影响。大数据应用也将对社会的组织结构、经济运行机制、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活、工作和思维方式等产生深远的影响。

本书由两大部分组成，第一部分介绍大数据管理理论框架和生态系统，包括大数据概述；大数据战略和商业模式变革；大数据平台的架构体系；大数据的数据整合、交换与交易；大数据管理和治理；最后提出大数据创新方法论。第二部分介绍数据科学和数据工程，包括数据科学理论和工具；医疗健康大数据解决方案、环保行业大数据解决方案、移动社交行业大数据解决方案、金融大数据解决方案、中国制造大数据解决方案和大数据工程保障体系建设。

大数据是综合性较高的交叉学科，本书全面、系统地阐述了大数据管理和技术、大数据科学和工程，具有很强的理论指导性和实践意义。本书可以供企业管理者、数据科学研究工作者、首席信息官等作为参考资料，也可以作为企业管理、计算机、软件工程等相关专业学生的教材使用。

图书在版编目(CIP)数据

大数据·数据管理与数据工程/赵眸光，赵勇编著.—北京：清华大学出版社，2017
ISBN 978-7-302-46928-5

Ⅰ. ①大… Ⅱ. ①赵… ②赵 Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 074347 号

责任编辑：郑寅堃 薛 阳
封面设计：
责任校对：焦丽丽
责任印制：

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 **邮 编：**100084
社 总 机：010-62770175 **邮 购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：
装 订 者：
经 销：全国新华书店
开 本：185mm×260mm **印 张：**26.25 **字 数：**632 千字
版 次：2017 年 6 月第 1 版 **印 次：**2017 年 6 月第 1 次印刷
印 数：1～ 000
定 价： .00 元

产品编号：070944-01



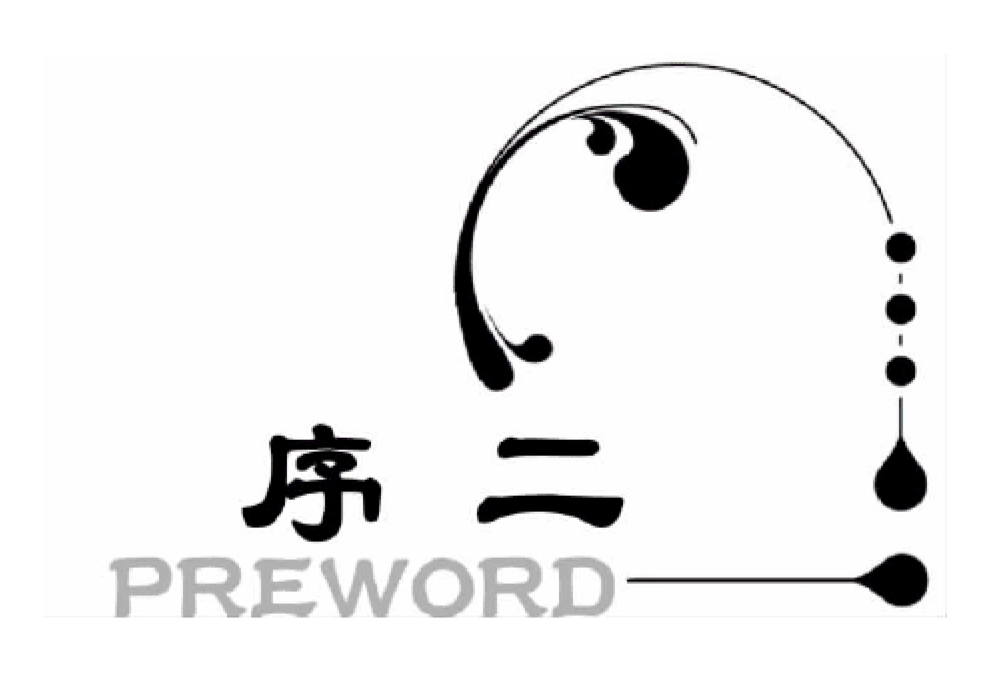
大数据是信息领域的前沿技术。大数据时代的来临，使人类有可能在浩如烟海的技术领域中，通过使用各种数据，发现和探索自然世界的规律。大数据时代的物理科学、计算科学、生命科学、社会科学及其他许多科学门类都将发生本质上的变化和发展，进而对人类的生产方式、生活方式和学习方式产生深刻的影响。

信息技术与经济社会的交汇融合引发了数据的迅猛增长，大数据已成为国家基础性战略资源，发展大数据及其相关技术研究更是重塑国家竞争优势的新机遇。国务院在2015年印发的《促进大数据发展行动纲要》中就强调了发展新兴产业大数据和工业大数据管理应用的重要性。通过大数据这种创新方式来解决我国在教育、交通、医疗和工程行业现代化所面临的种种问题，创建新的产业群，对实现由“中国制造”到“中国智造”再到“中国创造”有着重大意义。

云计算、物联网、移动互联网等新兴服务促使人类社会的数据种类和规模正以前所未有的速度增长。大数据具有Volume(大量)、Velocity(高速)、Variety(多样)、Value(低价值密度)、Veracity(真实性)这“5V”特性。对制造企业而言，大数据技术的战略意义不仅在于掌握庞大的数据信息，更在于对数据的“加工能力”，即对大量数据进行专业化处理的能力，使之转化成为对企业有价值的信息。制造企业如果能够在工业环境中建立起大数据平台，提高工厂对不同设备收集的海量信息进行数据挖掘的能力，提高企业信息系统的计算能力和数据处理能力，实现对企业的产品数据、运营数据、销售数据、客户数据的实时而有针对性的分析，用于洞察市场先机、客户需求，优化生产与管理流程，降低成本、提高运营效率、实现精准营销等，使得企业能够在成本有效控制的条件下，实现智能化生产、协同化组织和个性化服务。

赵眸光和赵勇博士多年在大数据理论、技术与应用等方面深入研究，取得了一系列成就。本书重点围绕大数据管理和大数据工程两方面进行了系统化的阐述，研究了大数据平台的体系架构和数据整合、交换与交易技术，通过对大数据的管理，总结出大数据创新方法论。此外，本书详细介绍了数据科学理论与工具，包括数据仓库、数据挖掘和知识发现等，对于医疗行业、移动社交、工业制造等几个热点行业数据工程的实践，进行了有针对性的阐述。全书内容系统，论述充分，为高校、研究院所科技研究人员和企业工程技术人员、管理人员从事大数据研究、应用和培训提供了一本极好的参考书。特此推荐。

中国工程院　院士
中国大数据技术与应用联盟　理事长
浙江大学　教授
2017年4月



数据自古存在。

乌龟壳、树皮、绸缎、竹简都曾是记录数据的媒介，留声机、磁带机也曾经风靡过，就连现在的信息技术，像个人电脑、智能手机、iPad在不远的将来也将会退出舞台，唯有数据，虽然不断地变换表现形态，却将一直伴随人类走向未来。

物联网本质上是器物层面的技术，从大数据的视角而言，是采集数据的终端。云计算本质上是传统计算机和网络技术发展融合的产物。物联网和云计算都是信息技术发展的一定阶段的自然延伸，依然属于信息技术范畴。而大数据其实是传统数据发生的质变。大数据超越信息技术，使人们重新界定国家竞争的主战场，重新审视政府治理水平，重新认识科学研究的新范式，重新审视产业变迁的驱动因素，重新理解投资的决策依据，重新思考公司的战略和组织。总之，大数据是推动经济发展、保障国家安全和社会治理的永恒主题。

大数据蕴含巨大价值，是国家意志和主权不可分割的部分。

2012年3月，奥巴马发布美国版的《大数据发展计划》时，我曾经写过一段点评："国家层面大数据技术领域的竞争事关一国的安全和未来。国家数字主权体现为对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。"在这篇点评中，明确提出数字主权的概念，坦言大数据必须上升为国家意志，落实为国家战略。2014年5月1日，美国白宫发布了《美国白宫：2014年全球"大数据"》白皮书，阐述了大数据带来的机遇与挑战。2014年8月，联合国开发计划署首次携手科技企业共建大数据实验室。我国2015年9月《促进大数据发展行动纲要》出台，赋予了大数据作为建设数据强国、提升政府治理能力和推动经济转型升级的战略地位。

保护国家层面的数据安全，恰恰是以数据开放为基础的。开放是一种态度，更是一项能力。一些重大基础数据开放，可以构成社会的数据基础，按照大数据定律之一"数据之和的价值远远大于数据价值的和"来推断，来自不同领域的数据聚合在一起，开放给社会，将会产生类似核聚变一样的价值发现效应。

开放的数据是基础，促使信息产业繁荣，才能诞生真正的数据驱动的企业，这些企业反过来在数据领域的技术进步，才是确保国家数据安全的长治久安之策。很难想象，如果没有谷歌、亚马逊、Facebook、苹果这样的公司，单凭美国政府一己之力能够实施如此庞大的"棱镜"计划吗？所以制定国家大数据战略，需要重新思考传统的所谓的"国家机密"和国家安全的关系。应当把消除部门数据格局，建立公开、透明、共享的数据公共平台作为长期的战略目标。

大数据将成为政府治理、企业管理、产业价值发现的重要工具。

大数据将打开各行各业的数据宝藏。政府治理、社交网络、医疗、教育、环保、金融、智能制造等，都会受益于大数据而被挖掘出更多的价值。在政府治理领域，通过让海量、动态、多

样的数据有效集成为有价值的信息资源，推动政府转变管理理念和治理模式，进而加快治理体系和治理能力现代化。还有推动政府治理决策精细化和科学化。如何将海量数据对行业进行管理决策、产品设计、精准营销、客户个性化服务等？如何对行业大数据进行商业模式设计？如何进行大数据平台建设？如何发挥大数据性能优势？如何解决安全和隐私？如何对各种大数据进行可视化？达到好的效果？本书解决了大数据从顶层设计到应用落地，从商业模式到技术平台，从数据管理到数据价值发现。本书的出版，正如天降甘露，恰到时机。

数据科学——科学地研究数据，用数据来研究科学。

大数据的五大特征(体量、类型、速度、价值和真实)蕴含了大数据丰富的内涵和外延。学术界在大数据时代有了广阔的舞台。大数据的早期发展是由技术性公司推动起来的，例如谷歌、亚马逊等一线互联网公司。产学研合作也正好是推进大数据发展的最佳途径。学术界有很好的理论基础和算法优势，产业界有很好的支持平台。鄂维南院士呼吁学术界向谷歌公司学习，同时指出："大数据在科学领域的表现是数据科学的兴起，数据科学将成为科研体系中的重要组成部分"，也为数据科学发展指明了方向。

在大数据时代，许多学科表面上看来研究的方向大不相同，但是从数据的视角来看，其实是相通的。比方说自然语言处理和生物大分子模型里之所以都用到隐式马氏过程和动态规划方法，其最根本的原因是它们处理的都是一维的随机信号；再如用于图像处理的算法和用于压缩感知的算法也有着许多共同之处。

图灵奖得主格雷(Jim Gray)提出科学研究的"第四范式"是数据，不同于实验、理论和计算这三种范式。在该范式下，需要"将计算用于数据，而非将数据用于计算"。吴军博士在《数学之美》一书中也讲到了这方面的故事。以自然语言的机器翻译研究为例，最初科学家们都是试图为计算机建立一系列的语法规则，按照语法、词义来翻译成另外一门语言。这个思路非常直观，因为人们就是如此理解学习的语言的。但是在实践中却困难重重，基于语法规则的翻译器，几乎就没有商用过。而当科学家们改弦易张，计算每一个词、每一句话的"合理概率"时，复杂的机器翻译就简化成了文字的概率计算。通俗地说就是："如果大多数人都这么说，就认为是对的。"这种思想在越来越多的领域得到应用。比如宏观尺度研究的天体信息学、社会行为学，微观尺度上分析人类的基因组、追踪物理学家们梦寐以求的"上帝粒子"。

随着大数据应用领域的逐步深入，越来越多的应用在数据层面趋于一致。数据科学在数学、概率模型、统计学等和实际应用之间建立起了直接的桥梁。本书在数据科学理论方面建立起了有效的方法论体系。

数据工程——大数据产业发展支撑体系。

曾经和中关村大数据产业联盟几位专家、总裁一起讨论，大家七嘴八舌地提出"＋大数据"的概念。希望在联盟中培育出各个专家组，把大数据思维嫁接到不同的产业，推动大数据在各行各业落地。大数据产业变革综合运用了大数据相关理论。本书介绍了医疗、环保、社区、金融和智能制造大数据产业分析和系统架构实现，对其他行业的应用也有很好的指导作用。许多行业龙头也开始蠢蠢欲动，应用大数据思维解决产业变革问题。例如农业领域的大北农、教育行业的新东方、玩具领域的奥飞动漫……

给企业家们带来冲击的不仅仅是大数据引起的产业变革，更是一些新兴公司的不可思议的跨界能力。就像本书中指出的那样，行业之间的界限变得越来越模糊，这些新兴的"野

蛮人”采用新的技术、新的模式，大规模采集数据，迅速形成预判，然后就以看似“野蛮”的方式扩张到其他行业。譬如卖农产品的去搞金融服务，做金融业务的帮助企业做采购等等，不一而足。

传统产业的各行各业，都面临在大数据和移动互联网时代如何彻底转型和再造问题。产业整合，也在大数据时代出现了全新的整合逻辑和实现契机……我仿佛看到了一个未来景象：传统产业都可能在大数据和移动互联时代重现生机、焕发青春。当然，与此对应的是，凡是不能跟上这个时代步伐的企业和行业，将会退出历史舞台。

在星空格局之下，公司的竞争力更多体现在“平台＋特种部队”模式。就像美军前线的一个小分队，甚至单兵可以直接指挥后方的导弹、飞机一样。以星空格局作为产业演化的最终形态，以特种部队作为业务竞争的基本单元，整个公司的战略、组织、文化等方面需要彻底的重组。传统公司的确需要重新审视自己的战略，重构组织，再育文化，这也是大数据思维非常重要的原因。

综上所述，不能狭隘地看待大数据，不能将其作为数据挖掘的工具，不能唯技术论。很欣慰看到两位学者编写的《大数据·数据管理与数据工程》一书，不是就技术而谈技术，而是从更宽广的视角阐释大数据带来的冲击、管理理念的变革以及大数据生态系统。尤为重要的是，提出数据工程的概念，奠定了大数据应用领域标准化、工程化的基础。

从大历史观来看，“大数据”的内涵远远超越物联网、云计算等信息技术的概念，它的意义可以比肩活字印刷术的发明。“大数据”将在世界尺度上大范围地消除信息不对称的现象，释放巨大的生产力，深刻改变社会的面貌，革新科学研究的思想，促进产业间的跨界、融合和颠覆，并将极大地促进文明的传播、凝聚和升华。

是以为序！

中关村大数据产业联盟秘书长

2016 年 12 月



序三

PREWORD

建立互联网金融治理体系，应该成为我国金融治理体系和金融治理能力建设的重要内容，大力发展互联网金融，以互联网金融治理推进中国金融治理体系和治理能力现代化，是金融治理创新发展的重要引擎。凯文·凯利(Kevin Kelly)被誉为互联网经济的预言家，他精准预测 Web 2.0 时代的到来和网络经济的运行规律。凯文·凯利预言，未来，大数据、云计算、移动通信三者相结合的技术进步将激发大数据、深度学习、人工智能、P2P、虚拟货币等方面的技术突变，而这些正在成为现实。未来技术改变的世界有四大特征：万物互联、信息交互、数据集成、智能决策，这四大特征正是物联网大数据时代的主要特征，这也正是金融模式创新的基础前提。

从该书大数据市场行业应用分析中可以看出，金融行业在大数据应用可行性和市场成熟度方面都属于优先级比较高的领域，是大数据应用热点。本书提纲挈领、高屋建瓴地从大数据系统科学的角度去认识"大数据"，指出大数据的内涵和研究方向，从而发现大数据的价值，通过全球大数据的战略视角，窥视行业应用商业模式和商业机会。从金融创新来看，数据成为资产、行业垂直整合、平台泛金融化成为商业发展主流趋势，行业产业链条加深加长，促使商业创新模式层出不穷。互联网创造出新的商业模式，塑造新的经济形态，创造新的经济生态空间，加大生产可能性边界，降低生产成本和融资成本，互联网基因已经融入到社会运行的底层物质技术结构之中。大数据时代的金融创新，必将发生像作者在书中提到的种种金融变革。

该书从大数据架构体系、安全和隐私、系统整合、数据管理以及理论创新方面全面系统地提出管理方法和技术工具，通过数据科学理论在金融创新和风险控制方面，在大数据征信贷款、大数据反欺诈、大数据客户管理和精准营销方面做出了分析。例如大数据技术运用于信贷技术前，借款需要很长时间的审核，尤其是线下取证、财务报表、抵押担保、审批流程、领导签批、最后借款等环节。根据内在的大数据信用评估和内控技术，能够实现实时计算借款人的信用额度，在信用额度内实现即时放款。这在传统金融领域是难以想象的，而这种快速借款模式，将成为未来互联网金融时代的标志。

该书体系完整、结构清晰、逻辑严谨，是大数据从战略到战术、理论到实践、产业到模式、标准到工程，具有战略性、系统性、理论性和指导性的大数据百宝箱和重要参考全书。当前，国家大数据战略日渐清晰，产业应用初具规模，大数据技术日趋成熟，本书为大数据从业者和应用机构提供了大数据应用知识地图、全新的认识和决策思路，非常值得一读。

大数据金融创新的数据可视化已经成为经济分析、管理决策、绩效评价等工作的重要工具。金融可视化是利用数学模型、网络技术、数据挖掘、计算机语言等一系列数据科学前沿科技综合应用的重要成果。该书提供了丰富的金融数据可视化展示工具和方法，不仅能够让数据丰富多彩地展示，还原真实世界，得出精准信息，更让人们能够通过数据模型直观地

感受到数据的真实变化。数据使得决策更加科学化、智能化、动态化、实时化，成为决策的重要依据。

从金融业的发展趋势来看，大数据技术将会成为风险管理的最佳工具，云计算为金融业务的高效实时处理做出保障，点对点的资源配置方式充分发挥金融职能，越来越多的传统金融需要这些互联网金融新模式作为技术载体、信息载体和业务载体。互联网金融对现代金融业的塑造主要体现在互联网金融平台上，通过自我创造、自我发展衍生出金融业务交易平台、新兴技术应用平台、风险控制管理平台、金融模式创新平台和普惠金融服务平台。本书在数据工程实现和金融平台建设上提供了技术支持保障。

书中在大数据管理创新和工程实践中提供了全新的视角和系统性思维，在目前大数据领域丛书中，具有更强的指导性。随着应用的不断深入，学习和研究也要与时俱进。互联网金融会成为金融创新发展的必然趋势。新的技术不断涌现、智能搜索引擎、区域链技术、全新的信息通信和物联网技术等必将会对金融业产生革命性的影响，也为互联网金融的发展提供一个良好的契机，可以让金融监管发挥更大的效力。先进的大数据金融信息系统可以及时检测金融市场与企业的动态，而电子化的渠道可有效地降低监管的搜索成本，多渠道的信息数据来源可以降低监管面对的信息不对称难题，而通过机器学习可以构建智能监管监测系统。这些信息化金融监管手段来源于市场，作用于市场，检测于市场。金融是现代经济的核心，推进我国互联网金融治理体系和治理能力现代化，是金融治理创新和经济发展的必由之路。本书一定会成为大数据青睐者和行业践行者的良师益友。

姚余栋

中国人民银行金融研究所所长



信息作为一种资源自古就存在，信息就是物质，信息通过电子化、数字化无限增值。1800年，伏特发明了世界上第一块电池；1946年，人类发明第一台电脑。伴随电脑、互联网时代的到来，信息成为可生产、交换、传播的商品。个人电脑、互联网、浏览器、搜索引擎、智能手机、社交网络、可穿戴设备、3D打印比过去基于蔡伦、毕升、古登堡时代，传统印刷更为丰富、多元、有效。不到半个世纪，人类存储的数据量以指数级在增长，数据传输速度从数天缩短到数毫秒，提升达9个数量级，成为全球拥有、共享、传播的大数据海量信息。随着全球大数据、物联网、云计算、移动社交网络等信息网络新技术的普及，推动世界数字经济呈指数增长，人类社会信息化进入大数据时代。

然而，数据规模如此之大、数据结构如此复杂、数据传播如此之快，已经远远超过了目前政府或企业在数据采集、存储、处理和分析、管理和应用方面的能力。企业如何发现数据的价值？如何利用数据产生效益？大多数企业还是手足无措。

本书通过大数据管理理论框架与生态系统、数据科学与数据工程两大部分，基本上覆盖了数据起源、数据架构（基础设施、数据采集、存储、分析处理、可视化、应用、运维、安全和隐私）、数据整合与交换及交易、大数据管理与治理、数据创新与数据科学、重点行业应用等。全面解决了大数据如何应用和价值发现的过程。

大数据成为全球重要的战略资源和核心资产。大数据时代，各国对数据的依赖快速上升，国家竞争焦点已经从资本、土地、人口、资源的争夺转向了对大数据的争夺，对大数据的开发、利用与保护的竞争日趋激烈，制数权成为继制陆权、制海权、制空权之后的新制权。大数据使得强国与弱国不再以经济规模和经济实力论英雄，而是取决于一国大数据能力的优劣。

借助大数据革命，美国等发达国家全球数据监控能力升级，美国先后推出《网络空间国际战略》《网络空间国际行动》等重要战略规划，确保自身在网络和数据空间的主导地位。

中共中央十八届五中全会提出，要拓展发展新空间，实施网络强国战略，实施“互联网＋”行动计划，发展分享经济，实施国家大数据战略。国务院通过《关于促进大数据发展的行动纲要》为未来中国的大数据发展指明了方向。

据统计，2015年全球信息社会指数为0.5494，正在从工业社会向信息社会加速转型，专家预计人类2018年进入信息社会。中国互联网经济占GDP比重4.4％，已超过美国、法国和德国，达到全球领先国家水平。要实现两个百年发展目标，2021年中国人均信息消费将接近1000美元，2049年中国人均信息消费将超过3000美元，成为世界最大的信息经济体。2013年中国大数据产业市场规模为34.3亿元，同比增长率超100％，未来一段时间将持续快速增长。2014年7月，麦肯锡全球研究员发布的《中国的数字化转型：互联网对生产力与增长的影响》预测：2013到2025年，互联网将占到中国经济年增长率的0.3％～1.0％，互

联网将可能在中国 GDP 增长总量中贡献 7％～22％，我国正从数据大国向数据强国过渡。

中国作为世界最大的发展中国家，能否吸取工业革命中“落后挨打”的悲剧教训，在全球化信息网络时代跨越中等收入国家陷阱和修昔底德陷阱？中国能否在这次全球信息革命浪潮中抢占先机、立于不败之地？能否实现中华民族伟大复兴的中国梦、两个百年目标？

我国必须要紧抓大数据技术发展机遇，正如本书所述，建立起大数据标准体系、数据科学理论体系、标准化大数据治理体系，实现弯道超车快速崛起，成为全球最大信息经济体的受惠者。

伍刚

中央人民广播电台高级编辑、央广网副总裁



大数据历经几年的发展,在全球已进入了高速发展期。我国“十三五”规划正式将大数据上升为国家战略,当前全国各省市级和地区级城市正在制定大数据发展战略和实施规划,中国正在创造一个万亿级的大数据市场。在此期间,笔者 2014 年编著了《大数据革命——理论、模式与技术创新》,2015 年又出版了大数据的技术教材《架构大数据——大数据技术与算法解析》。在大数据产业发展上,以成都为基地,成立大数据协会和联盟,如四川大数据产业联盟、中国西部互联网与大数据产业协会等,提供大数据人才培训和培养、政府大数据产业规划和企业转型升级咨询。成立第五维国际大数据孵化器,通过和硅谷孵化器合作,为大数据创业团队提供导师、技术、办公场地和资金等全方位的孵化服务。在大数据产品研发上,以清数科技公司为依托,开发了 Neo 大数据一体“傻瓜机”,把数据从采集、存储、处理、分析和挖掘、可视化和应用服务全部集成部署到一体机服务器中,让政府和企业拥有“开箱即得”的大数据分析处理能力,方便了用户的操作使用。

本书正是笔者在大数据产品研发和产业落地基础上的理论升华和管理思考。笔者预测,中国的大数据产业将在明年中期迎来应用的全面爆发,大数据的平台、分析、应用类的产品和服务将供不应求。而大数据交换和交易的市场,随着国务院制定的政府数据开放日程的临近(《大数据发展行动纲要》要求各部委数据在 2018 年底完成开放),也将在两年后成为大数据产业的最大的市场,数据资产、数据产品、数据服务都会带来巨额的财富。本书正是顺应大数据发展趋势,重点阐述了大数据生态系统、大数据管理、数据交换、共享、交易等理论体系,数据科学理论和大数据行业应用实践,以及相应的大数据标准体系;全面系统地阐述了大数据体系建设和工程实践,真正挖掘和实现了大数据的价值。本书内容主要围绕大数据应用热点和重点行业展开分析,如医疗、环保、社交、金融、工业制造等,这些理论实践同时也适用于教育、政务、交通、能源、航空、农业、旅游等行业的发展应用。总结出了大数据管理创新方法论和工程实践经验,为中国大数据产业发展和创新生态链打造奠定了理论和实践基础。

众所周知,从上届美国总统的选举到本届美国总统选举,无疑都是大数据应用的最好例证。本届选举演变成了希拉里和特朗普背后的大数据团队的生死角力。双方都拥有阵容强大的大数据团队,服务于特朗普的 Deep Root Analytics(深根分析)公司和英国的剑桥分析公司采取的是类似于精准广告投放的技术,分析摇摆投票者们的意识形态、价值观以及他们喜欢的信息接收方式和渠道,然后针对他们制定竞选演说、拉票方式和信息传递方式,最终帮助特朗普问鼎美国总统宝座。尽管是在被业界称为投资寒冬的大环境下,大数据以及人工智能还是在美国硅谷和中国的投资圈刮起一股旋风,数百家相关的大数据企业都顺利拿到了投资。大数据应用成为产业聚焦的热点。

本书的编写得到了很多协会和清数的同事们的支持和帮助,尤其是李小龙、张晓东、唐

犀、赵虎、滕雨橦，还有电子科技大学极限网络计算与服务实验室的同学们，他们为本书收集了大量的资料，并提供了很多的内容。我也要感谢我的家人们对我的鼓励和支持，很多节假日都没能陪同她们。

本书由于笔者的知识和经验有限，存在的疏漏敬请读者原谅，也欢迎与我们联系，我们一起为中国的大数据事业贡献力量，谢谢大家。

赵　勇

2016.12



大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的必然产物，也必将对网络通信(ICT)和物联网(IOT)产业产生深远的影响。大数据技术的发展与应用，将对社会的组织结构、经济运行机制、社会生活方式、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活、工作和思维方式等产生深远的影响。随着社会网络安全、应急管理、医疗健康、经济金融、交通运输、制造领域、社交社区等各个领域大量数据的使用，对于我们而言，能够及时、有效地了解数据和信息的意义，进而改善决策制定的过程将变得尤为重要。大数据的价值必将对现代企业的管理运作理念、市场营销决策以及消费者行为模式等产生巨大影响，使得企业商务管理决策越来越依赖于数据分析而非经验甚至直觉。因而，大数据也必将对这种传统的商业模式进行近乎彻底的颠覆与模式的重构。

当前，美国、日本、法国、韩国、澳大利亚等国家相继启动了推动大数据产业发展的政策改革，并把大数据产业发展纳入国家发展战略，通过有力的资金和政策支持加强大数据研究，优化其发展环境，抢占大数据产业发展的制高点，使其成为推动国民经济社会发展的新手段。鉴于发达国家对大数据产业的强力推动，大数据在经济、国家安全、社会、科研等方面的巨大价值和适应经济社会发展的要求，中国各级政府和社会各界也纷纷制定相关政策推动大数据产业深入发展，运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势，我国相继制定实施大数据战略性文件，大力推动大数据发展和应用。目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，一些地方政府已启动大数据相关工作。坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

我们认为大数据的发展必将经历三个重要的阶段。①“技术驱动型”。大数据的核心关键技术正在加速发展和快速迭代，技术体系框架也已日趋成熟，基本能够满足产业发展需求，比如 Hadoop 生态框架系统。大数据架构体系分为基础设施、采集、存储、处理、分析、应用、安全和维护几个方面。②“行业驱动型”。各大解决方案服务商围绕电信、环保、金融、交通、医疗、政府、教育、工业、城市管理、社交网络等重点行业领域描绘美好蓝图，力求推动行业应用，如节能环保产业布局了高效储能、节能监测和能源计量；生物医药产业布局了生物资源样本库、基因测序，以及基于物联网的远程健康管理服务等。这一阶段发展虽然还有距离，但这一转变过程正在加速进行。③“模式驱动型”。大数据行业应用深化发展，使得领域和行业边界愈加模糊，商业模式应用创新超越技术本身，企业以独特数据资源进行的整合朝着纵向产业链上下游整合和横向多种产业整合两个方向发展，生产模式向服务化转变，数据作为一种资产资源为企业带来新的商业价值，数据开放为政府治理和个人福祉都带来新的机遇。

从大数据系统论的角度，可以将大数据划分为大数据技术、大数据管理、大数据科学和大数据工程，本书重点围绕大数据管理和大数据工程两部分展开阐述。

第一部分介绍大数据管理理论框架和生态系统，共分为6章，主要内容有：数据时代背景、大数据定义、特征、数据结构、度量价值、数据管理与技术和大数据科学与工程研究方向以及大数据生态系统；国内外大数据战略和大数据应用的商业模式变革；大数据平台架构体系自下而上包括基础设施、数据采集、数据存储、数据处理、数据可视化、大数据应用、运维和数据安全；大数据平台整合、大数据与存储、大数据与网络、大数据与虚拟化技术整合、大数据环境的数据整合、大数据交换和数据交易；大数据流程管理、大数据事务管理、大数据技术管理以及大数据质量管理阐述；最后提出大数据创新理论指标体系、大数据创新重要环节和大数据创新最佳实践。该部分章节框架清晰、结构分明、逻辑严谨、层次有序、概念明确、重点突出、体系完整，形成整个大数据技术管理体系。

第二部分介绍数据科学和数据工程内容，共分为7章，主要内容有：数据科学概念、研究重要角色、生命周期管理、数据仓库、数据挖掘分析方法、知识发现及大数据处理平台，通过建立科学系统的数据分析方法论，指导数据工程实践；在数据工程方面，重点介绍医疗行业大数据、环保行业大数据、移动社交大数据、金融行业大数据和工业制造大数据等几个热点行业数据工程实践，每个行业又侧重大数据应用的不同角度，总体上全面解析大数据应用的多个方面；医疗健康主要包括总体架构(业务架构、技术架构和网络架构)、医疗大数据存储处理、容灾备份解决方案和医疗大数据分析等；环保行业包括环保物联网架构、电力脱硫工作原理、电力脱硫数据分析优化目标以及空气质量大数据分析评价体系；移动社交包括发展趋势、社交理论、社交网络商业模式、社交网络平台以及社交网络数据分析；金融行业包括金融大数据特征、发展机会、总体架构(业务架构、技术架构和网络架构)、金融大数据风险管理平台、大数据征信、大数据反欺诈、大数据精准营销以及大数据带来的产业变革；工业大数据通过回顾全球工业信息化发展历程和现状，提出了中国制造2025发展战略，同时指出工业信息技术集成和协同发展方向，利用工业信息化应用系统搭建工业大数据架构体系(业务架构、技术架构和安全架构)、智能化协同制造架构原理，最终实现智能化协同制造服务。工业是国民经济的基础，工业的未来也是我国经济发展的未来。最后提出大数据工程保障体系建设，包括法律体系建设、标准体系建设、标准化大数据治理体系建设、技术和应用研究、创新平台建设等，该部分章节充分体现了理论性、科学性、创新性、实用性、经济性、社会性、标准性、保障性和完整性，形成了数据科学和数据工程体系。

本书是作者和在大数据研究领域非常有名望的赵勇博士共同编写而成的。书中的第3～6章来源于赵勇博士研究成果，其他是作者多年来对物联网、云计算和大数据的研究、咨询和应用实践经验的智慧结晶，同时也是在清华大学继续教育学院致力于智慧城市规划设计和企业管理咨询工作经验的积累。希望本书将我们多年从事于大数据研究方面的成果展现给读者，本书可以作为企业管理者、数据科学研究工作者、首席信息官等的参考资料，也可以作为企业管理、计算机、软件工程等相关专业学生教材使用。

本书在撰写的过程中，得到了清华大学、北京大学多位老师，清华大学数据研究院和行业同仁的资料提供和支持帮助，在此表示衷心的感谢！也感谢我的家人给予我莫大的支持和鼓励，使我顺利完成写作。大数据发展日新月异，相关技术快速发展，由于我们对大数据的理解和知识水平都有局限，书中疏漏或不足之处在所难免，敬请读者批评指正。

赵眸光

2016年12月于清华园



第一部分　大数据管理理论框架与生态系统

第 1 章　大数据概述 …… 3

1.1　大数据时代 …… 3
1.2　什么是大数据 …… 4
1.2.1　大数据定义 …… 4
1.2.2　大数据特征 …… 5
1.2.3　大数据结构类型 …… 5
1.2.4　数据、信息、知识与智能的关系 …… 6
1.3　大数据发展史 …… 9
1.3.1　数据管理发展历程 …… 9
1.3.2　大数据的演变及回顾 …… 12
1.4　大数据的度量和价值 …… 15
1.4.1　大数据的度量 …… 15
1.4.2　大数据的价值 …… 15
1.5　大数据生态系统 …… 17
1.5.1　大数据生态系统全貌 …… 17
1.5.2　大数据生态系统框架 …… 18
1.6　大数据应用研究方向 …… 21
1.6.1　大数据管理与技术 …… 22
1.6.2　大数据科学与工程 …… 22
1.7　大数据的挑战 …… 23
1.7.1　大数据管理方面带来的挑战 …… 23
1.7.2　大数据技术方面带来的挑战 …… 23
1.7.3　大数据工程方面带来的挑战 …… 23

第 2 章　大数据战略与商业模式变革 …… 25

2.1　大数据战略 …… 25
2.1.1　国外大数据战略视角 …… 26
2.1.2　国内大数据战略视角 …… 29

2.2 大数据商业模式和商业机会 …… 32
2.2.1 基于大数据的商业模式创新 …… 32
2.2.2 大数据对企业管理决策的影响 …… 38
2.2.3 基于大数据驱动的商业机会 …… 39
2.3 大数据市场的行业应用需求 …… 44
2.3.1 移动互联网和社交网络 …… 44
2.3.2 政府公共管理 …… 46
2.3.3 教育科研行业 …… 48
2.3.4 金融行业 …… 50
2.3.5 医疗健康业 …… 51
2.3.6 中国制造 2025 …… 52
2.3.7 智能交通领域 …… 54

第 3 章 大数据平台的架构体系 …… 56

3.1 大数据基础设施 …… 56
3.1.1 虚拟化 …… 57
3.1.2 云计算 …… 57
3.1.3 数据中心 …… 62
3.2 数据采集 …… 63
3.2.1 系统日志采集方法 …… 63
3.2.2 网络数据采集方法：对非结构化数据的采集 …… 63
3.2.3 其他数据采集方法 …… 63
3.3 数据存储 …… 67
3.3.1 结构化数据存储 …… 69
3.3.2 非结构化数据存储 …… 70
3.4 数据处理 …… 71
3.4.1 离线批处理 …… 72
3.4.2 实时交互计算 …… 74
3.4.3 流计算 …… 76
3.5 数据交互展示 …… 78
3.5.1 数据可视化基础 …… 79
3.5.2 数据可视化模式 …… 80
3.5.3 数据可视化工具 …… 81
3.6 大数据应用 …… 84
3.7 运营管理 …… 85
3.8 安全管理 …… 85

第4章 大数据的数据整合、交换与交易…… 87

4.1 大数据平台整合…… 89
4.1.1 HDFS 分布式文件系统 …… 90
4.1.2 MapReduce 分布式计算框架 …… 91
4.1.3 HBase 分布式数据库 …… 94
4.1.4 交互式数据查询分析 …… 95
4.1.5 数据收集、转换工具…… 96
4.1.6 其他大数据平台 …… 96
4.2 大数据与存储架构的整合…… 98
4.2.1 传统存储架构 …… 98
4.2.2 集群存储的发展 …… 99
4.2.3 基于 HDFS 的集群存储 …… 100
4.2.4 固态硬盘对内存计算的支持…… 101
4.3 大数据与网络架构的发展 …… 103
4.4 大数据与虚拟化技术的整合 …… 105
4.5 Hadoop 环境下的数据整合 …… 107
4.5.1 Hadoop 计算环境下的数据整合问题 …… 107
4.5.2 数据库整合工具 Sqoop …… 108
4.5.3 Hadoop 平台内部数据整合工具 HCatalog …… 109
4.6 大数据数据交换 …… 110
4.6.1 数据集成技术…… 111
4.6.2 数据交换体系应用框架…… 113
4.6.3 数据交换关键技术…… 114
4.7 大数据交易 …… 116
4.7.1 大数据交易产业链…… 118
4.7.2 大数据交易业务模式分析…… 120
4.7.3 大数据交易发展趋势…… 122

第5章 大数据管理和治理…… 124

5.1 建立数据驱动的管理体系和架构 …… 126
5.1.1 建立数据管理组织和团队…… 126
5.1.2 建立数据管理规章和制度…… 127
5.2 大数据治理体系 …… 127
5.2.1 数据标准管理…… 128
5.2.2 数据质量管理…… 129
5.2.3 元数据管理…… 130

5.2.4 主数据管理…… 131
5.2.5 数据资产的全生命周期管理…… 131
5.3 大数据技术管理体系 …… 134
5.3.1 数据类型和结构…… 134
5.3.2 数据存储管理…… 135
5.3.3 数据仓库和商业智能…… 137
5.3.4 数据计算和处理…… 138
5.3.5 数据展示与交互…… 138
5.4 大数据事务管理 …… 138
5.4.1 事务的基本属性…… 139
5.4.2 大数据事务管理机制…… 140
5.5 大数据流程管理 …… 140
5.6 大数据易用性管理 …… 142
5.7 数据的安全管理 …… 142

第 6 章 大数据创新方法论…… 148

6.1 大数据的爆发 …… 148
6.2 大数据创新理论 …… 150
6.2.1 大数据的宏观性和微观性…… 150
6.2.2 大数据的生产要素性…… 151
6.2.3 大数据的基因特性…… 151
6.2.4 大数据的催化剂特性…… 152
6.2.5 大数据的活性和流动性…… 152
6.2.6 大数据的黑洞效应和核聚变效应…… 152
6.3 大数据创新方法论 …… 153
6.4 信息演变趋势 …… 154
6.5 大数据创新实践闭环 …… 155
6.6 中国创新创业大数据版图 …… 156
6.6.1 大数据时代的数据管理…… 157
6.6.2 大众创业万众创新的浪潮…… 157
6.6.3 中国创新创业大数据版图的推出…… 158
6.6.4 双创版图中的大数据管理挑战…… 160
6.6.5 双创版图中大数据技术的集中运用…… 161
6.6.6 双创大数据版图的意义…… 163

第二部分 数据科学和数据工程

第 7 章 数据科学理论与工具…… 167

7.1 数据科学理论基础 …… 167

7.1.1 数据科学概念…… 167
7.1.2 数据科学预测预警分析…… 168
7.1.3 商业智能与数据科学…… 169
7.2 数据科学研究的重要角色 …… 170
7.2.1 数据科学家…… 171
7.2.2 数据科学与工程相关角色…… 172
7.3 大数据生命周期管理方法论 …… 172
7.3.1 数据分析模型概述…… 173
7.3.2 数据分析模型流程框架…… 175
7.3.3 数据分析模型创新案例…… 175
7.3.4 数据分析工具…… 183
7.4 数据仓库理论 …… 187
7.4.1 数据仓库的主要特征…… 187
7.4.2 数据仓库建模…… 187
7.4.3 数据仓库设计…… 188
7.4.4 数据仓库建设方法论…… 189
7.4.5 数据仓库相关技术…… 190
7.4.6 DW、OLAP 与 DM 的关系 …… 192
7.5 数据挖掘高级理论 …… 193
7.5.1 聚类分析…… 193
7.5.2 关联分析…… 197
7.5.3 回归和分类分析…… 202
7.5.4 时序模型…… 212
7.5.5 结构优化…… 214
7.5.6 深度机器学习…… 216
7.6 大数据语义分析知识发现 …… 221
7.6.1 大数据知识发现过程…… 221
7.6.2 大数据知识发现技术框架…… 225
7.6.3 大数据知识发现专家系统…… 225
7.6.4 企业大数据知识管理框架…… 229
7.7 大数据分析处理平台 …… 230
7.7.1 结构化大数据处理架构…… 230
7.7.2 非结构化大数据处理架构…… 233
7.7.3 主流大数据分析平台…… 236

第 8 章 医疗健康大数据解决方案…… 242

8.1 医疗信息化 …… 244

8.1.1 美国医疗信息化发展情况…… 244
8.1.2 我国医疗信息化发展趋势…… 247
8.1.3 医疗健康大数据挑战和机遇…… 249
8.2 医疗健康大数据综述 …… 250
8.2.1 医疗健康大数据类型…… 251
8.2.2 临床服务数据…… 252
8.2.3 公共卫生调查和监测数据…… 252
8.2.4 医学研究性数据…… 252
8.2.5 个人健康数据…… 252
8.3 医疗健康大数据总体架构 …… 253
8.3.1 建设原则…… 253
8.3.2 建设目标…… 253
8.3.3 医疗健康大数据业务架构…… 254
8.3.4 医疗健康大数据技术架构…… 255
8.3.5 医疗健康大数据网络架构…… 256
8.4 医疗健康数据中心解决方案 …… 257
8.4.1 医疗数据中心架构设计方案…… 258
8.4.2 集中存储解决方案…… 259
8.4.3 PACS 数据存储方案 …… 262
8.4.4 容灾备份解决方案…… 267
8.5 医疗健康大数据分析 …… 268
8.5.1 医疗实体对象建模分析…… 269
8.5.2 医疗个人健康档案建模分析…… 269
8.5.3 相关数据特征对比分析…… 271
8.5.4 临床信息学大数据分析…… 272
8.5.5 医学文献研究知识发现…… 273
8.6 医疗健康大数据展望 …… 275

第 9 章 环保行业大数据解决方案…… 277

9.1 环保物联网 …… 278
9.1.1 物联网概念…… 278
9.1.2 物联网基本架构…… 279
9.1.3 环保物联网数据…… 281
9.2 环保电力脱硫 …… 281
9.2.1 火电脱硫的重要性…… 281
9.2.2 火电脱硫系统工作原理…… 281
9.2.3 火电脱硫相关数据…… 282

9.2.4 脱硫性能优化目标 …… 282

9.3 火电行业脱硫大数据分析 …… 283

9.3.1 主要理论和方法 …… 283

9.3.2 最优化脱硫可调参数 …… 284

9.3.3 最小化脱硫系统成本 …… 285

9.4 空气质量大数据分析评价体系 …… 285

9.4.1 基于熵权的模糊综合评价方法的原理 …… 286

9.4.2 综合评价指标选择与数据来源 …… 287

9.4.3 环境质量综合评价结果及分析 …… 287

第 10 章 移动社交大数据解决方案 …… 290

10.1 移动社交网络发展情况 …… 291

10.1.1 移动社交网络发展现状 …… 291

10.1.2 移动社交网络发展方向 …… 293

10.2 社交网络基础理论和商业模式 …… 294

10.2.1 社交网络相关理论 …… 294

10.2.2 社交化商业模式 …… 296

10.3 移动社交网络数据处理架构 …… 297

10.3.1 移动社交网络服务架构模型 …… 297

10.3.2 Facebook 应用案例 …… 298

10.4 移动社交网络大数据分析 …… 302

10.4.1 社交网络平台行为影响分析模型 …… 302

10.4.2 社交网络单平台内影响力分析 …… 303

10.4.3 社交网络多平台影响力分析 …… 305

第 11 章 金融大数据解决方案 …… 307

11.1 金融信息化 …… 307

11.1.1 全球金融信息化发展历程 …… 307

11.1.2 我国金融信息化发展趋势 …… 308

11.2 金融大数据综述 …… 309

11.2.1 金融大数据的特征 …… 309

11.2.2 金融大数据的机遇和挑战 …… 310

11.3 金融大数据平台总体架构 …… 311

11.3.1 建设原则和目标 …… 312

11.3.2 金融大数据业务架构 …… 313

11.3.3 金融大数据技术架构 …… 314

11.3.4 金融大数据网络架构 …… 316

11.4 金融大数据分析…… 316
11.4.1 银行风险管理状况分析…… 316
11.4.2 金融大数据风险管理云平台…… 318
11.4.3 大数据征信…… 320
11.4.4 大数据反欺诈…… 323
11.4.5 大数据精准营销…… 325
11.5 金融大数据带来的产业变革…… 327

第 12 章 中国制造大数据解决方案 …… 330

12.1 全球工业信息化发展历程和现状…… 330
12.1.1 美国工业信息化发展历程和现状…… 331
12.1.2 日本工业信息化发展历程和现状…… 333
12.1.3 德国工业信息化发展历程和现状…… 334
12.1.4 我国工业信息化发展历程和现状…… 337
12.1.5 我国《中国制造 2025》的发展战略 …… 338
12.2 工业信息化技术集成和协同发展方向…… 340
12.2.1 集成和协同的空间跨度…… 340
12.2.2 集成和协同的时间跨度…… 341
12.2.3 集成和协同的重点和对象…… 342
12.2.4 主要的集成和协同技术…… 343
12.3 中国制造信息化应用系统…… 343
12.3.1 工业设计自动化系统…… 343
12.3.2 制造控制自动化系统…… 346
12.3.3 制造执行系统…… 347
12.3.4 柔性制造系统…… 348
12.3.5 工业互联网与 CPS 系统 …… 349
12.3.6 ERP 信息系统 …… 351
12.4 工业大数据架构体系…… 353
12.4.1 互联网催生工业大数据…… 353
12.4.2 工业大数据内涵特征…… 354
12.4.3 工业大数据业务架构…… 355
12.4.4 工业大数据技术架构…… 357
12.4.5 工业大数据安全架构…… 358
12.5 智能化协同制造体系架构…… 359
12.5.1 智能化协同制造发展需求…… 359
12.5.2 智能化协同制造总体架构…… 360
12.5.3 智能化协同制造设计思想…… 362

12.5.4 智能化协同制造应用场景…… 367
12.6 智能化协同制造服务生命周期过程…… 367
12.6.1 制造资源服务集成与发现…… 368
12.6.2 制造服务资源访问策略…… 371
12.6.3 制造服务资源的优化与智能调度…… 371
12.6.4 智能化协同制造研究与自学习机制…… 375
12.7 工业大数据展望…… 377

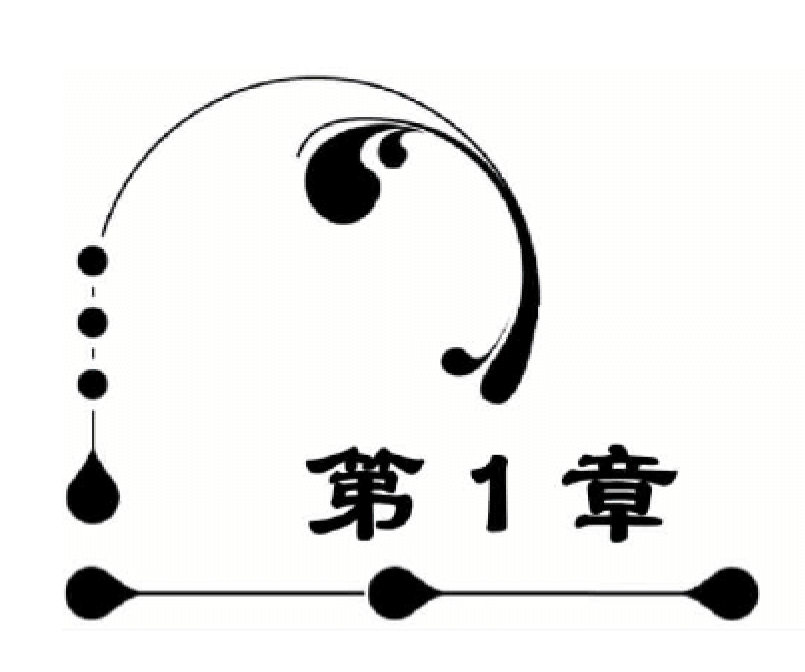
第13章 大数据工程保障体系建设 …… 378

13.1 法律体系建设…… 378
13.2 标准体系建设…… 380
13.3 建立标准化大数据治理体系…… 386
13.4 加强大数据行业应用研究…… 387
13.5 加强元数据的研究和应用…… 387
13.6 加强大数据核心技术研究…… 387
13.7 促进大数据交易市场的规范化发展…… 388
13.8 推动大数据标准化进程…… 388

参考文献…… 389

第一部分　大数据管理理论框架与生态系统

大数据管理理论框架和生态系统部分共分为6章，主要内容有：大数据时代背景、大数据定义、特征、数据结构、度量价值、数据管理和技术、大数据科学和工程研究方向以及大数据生态系统；国内外大数据战略和大数据应用的商业模式变革；大数据平台架构体系自下而上包括基础设施、数据采集、数据存储、数据处理、数据可视化、大数据应用、运维和数据安全；大数据平台整合、大数据与存储、大数据与网络、大数据与虚拟化技术整合、大数据环境的数据整合、大数据交换和数据交易；大数据流程管理、大数据事务管理、大数据技术管理以及大数据质量管理阐述；最后提出大数据创新理论指标体系、大数据创新重要环节和大数据创新最佳实践。



大数据概述

大数据(Big Data)已经成为人们耳熟能详的热点词汇,无论如何大数据都与人们每天发生着密切的关系。互联网、微信、邮件、微博、电话、导航、无所不在的监控和传感器等,无不表明我们已经进入物联网大数据时代。随着社会网络安全、应急管理、医疗健康、经济金融、交通运输、工业制造、社交社区等各个领域大量数据的使用,对于我们而言,能够及时有效地了解数据和信息的意义,来改善决策制定的过程将变得尤为重要。大数据是云计算、物联网、移动互联网、智慧城市等新技术、新模式发展的必然产物,也必将对网络通信(ICT)和物联网(IOT)产业产生深远的影响。大数据技术的发展与应用,将对社会的组织结构、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活、工作和思维方式等产生深远的影响。

1.1 大数据时代

随着以博客、社交网络、基于位置的服务等为代表的新型信息发布方式的不断涌现,以及云计算、物联网等技术的兴起,数据正以前所未有的速度在不断地增长和累积,大数据时代已经来到学术界、工业界甚至政府机构都已经开始密切关注大数据问题,并对其产生浓厚的兴趣。就学术界而言,*Nature* 早在 2008 年就推出了 Big Data 专刊;计算社区联盟(Computing Community Consortium)在 2008 年发表了报告 *Big data computing: Creating revolutionary breakthroughs in commerce, science, and society*,阐述了在数据驱动的研究背景下,解决大数据问题所需的技术以及面临的一些挑战。*Science* 在 2011 年 2 月推出专刊 *Dealing with Data*,主要围绕着科学研究中大数据的问题展开讨论,说明大数据对于科学研究的重要性。美国一些知名的数据管理领域的专家学者则从专业的研究角度出发,联合发布了一份白皮书 *Challenges and opportunities with Big Data*。该白皮书从学术的角度出发介绍了大数据的产生,分析了大数据的处理流程,并提出大数据所面临的若干挑战。

全球知名的咨询公司麦肯锡(McKinsey)于 2011 年 6 月发布了一份关于大数据的详尽报告 *Big data: The next frontier for innovation, competition, and productivity*,对大数据的影响、关键技术和应用领域等都进行了详尽的分析。2012 年以来,人们对大数据的关注度与日俱增。2012 年 1 月份的达沃斯世界经济论坛上,大数据是主题之一,该次会议还特别针对大数据发布了报告 *Big data, big impact: New possibilities for international development*,探讨了新的数据产生方式下,如何更好地利用数据来产生良好的社会效益。

该报告重点关注了个人产生的移动数据与其他数据的融合与利用。2012年3月份美国奥巴马政府发布了“大数据研究和发展倡议”(Big data research and development initiative),投资两亿以上美元,正式启动“大数据发展计划”。计划在科学研究、环境、生物医学等领域利用大数据技术进行突破。奥巴马政府的这一计划被视为美国政府继信息高速公路(Information Highway)计划之后在信息科学领域的又一重大举措。与此同时,联合国一个名为Global Pulse的倡议项目在2012年5月发布报告*Big data for development: Challenges & opportunities*,该报告主要阐述大数据时代各国特别是发展中国家在面临数据洪流的情况下所遇到的机遇与挑战,同时还对大数据的应用进行了初步的解读。《纽约时报》的文章*The age of big data*则通过主流媒体的宣传使普通民众开始意识到大数据的存在,以及了解大数据对于人们日常生活的影响。

我国正处于工业化向信息化发展的转型时期,信息的公开、共享与服务成为时代发展的主题。信息逐渐成为与物质和能源同等重要的资源,以开发和利用信息资源为目的的经济活动迅速扩大,逐渐占据或超越工业活动在国民经济活动中的地位。大数据的出现是跨学科技术与应用发展的结果。对于大数据,自然科学家强调在网络虚拟环境下对于密集型数据的研究方法,社会科学家则看重密集型数据后面隐藏的价值与推动社会发展的模式。

党中央、国务院高度重视大数据发展。党的十八届五中全会明确提出“十三五”时期要“拓展网络经济空间。实施‘互联网+’行动计划,发展物联网技术和应用,发展分享经济,促进互联网和经济社会融合发展。实施国家大数据战略,推进数据资源开放共享。”国务院《促进大数据发展行动纲要》(国发【2015】50号)明确指出要“建立标准规范体系。推进大数据产业标准体系建设,加快建立政府部门、事业单位等公共机构的数据标准和统计标准体系,推进数据采集、政府数据开放、指标口径、分类目录、交换接口、访问接口、数据质量、数据交易、技术产品、安全保密等关键共性标准的制定和实施。加快建立大数据市场交易标准体系。开展标准验证和应用试点示范,建立标准符合性评估体系,充分发挥标准在培育服务市场、提升服务能力、支撑行业管理等方面的作用。积极参与相关国际标准制定工作”。

1.2 什么是大数据

1.2.1 大数据定义

大数据本身是一个宽泛的概念,业界尚未给出统一的定义,不同的研究机构、公司从不同的角度诠释了什么是大数据。

2011年,美国著名的咨询公司麦肯锡(Mckinsey)在研究报告《大数据的下一个前沿:创新、竞争和生产力》中给出了大数据的定义:大数据是指大小超出了典型数据库软件工具收集、存储、管理和分析能力的数据集。根据Gartner的定义,大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

美国国家标准技术研究所(National Institute of Standards and Technology,NIST)的大数据工作组在《大数据:定义和分类》中认为:大数据是指那些传统数据架构无法有效处理的新数据集。因此,采用新的架构来高效率完成数据处理,这些数据集特征包括:容量、

数据类型的多样性、多个领域数据的差异性、数据的动态特征(速度或流动率,可变性)。

维基百科(Wikipedia)给出的定义是:大数据,或称巨量数据、海量数据、大资料,指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息。

百度百科给出的定义是:大数据,或称巨量资料,指的是所涉及的资料量规模巨大到无法通过目前主流软件工具,在合理时间内达到撷取、管理、处理并整理成为帮助企业经营决策更积极目的的资讯。

亚马逊网络服务(AWS)、大数据科学家John Rauser在2011年提到一个简单的大数据定义:任何超过了一台计算机处理能力的庞大数据量。

EMC公司给出的定义是:数据集或信息,它的规模、发布、位置在不同的孤岛上,或它的时间线要求客户部署新的架构来捕捉、存储、整合、管理和分析这些信息以便实现企业价值。

1.2.2 大数据特征

对大数据的完整理解应包含三个方面:数据特征、技术特征与应用特征。本书主要从大数据的数据特征来描述,业界通常引用国际数据公司IDC定义的4V来描述。

(1) 数据类型繁多(Variety):除了结构化数据外,大数据还包括各类非结构化数据,例如文本、音频、视频、点击流量、文件记录等,以及半结构化数据,例如电子邮件、办公处理文档等。

(2) 处理速度快(Velocity):通常具有时效性,企业只有把握好对数据流的掌控应用,才能最大化地挖掘利用大数据所潜藏的商业价值。

(3) 数据体量巨大(Volume):虽然对各大数据量的统计和预测结果并不完全相同,但是都一致认为数据量将急剧增长。

(4) 数据价值(Value):从海量价值密度低的数据中挖掘出具有高价值的数据。这一特性突出表现了大数据的本质是获取数据价值,关键在于商业价值,即如何有效利用好这些数据。

阿姆斯特丹大学的Yuri Demchenko等人提出了大数据体系架构框架的5V特征,如图1-1所示,它在上述4V的基础上,增加了真实性(Veracity)特征。真实性特征中包括可信性、真伪性、来源和信誉、有效性和可审计性子特征。

1.2.3 大数据结构类型

按照数据结构,数据分为结构化数据、半结构化数据和非结构化数据。结构化数据是存储在数据库里、可以用二维表结构来逻辑表达实现的数据。相对于结构化数据(即行数据,存储在数据库里,可以用二维表结构来逻辑表达实现的数据)而言,不方便用数据库二维逻辑表来表现的数据即称为非结构化数据,包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等。

所谓半结构化数据,就是介于完全结构化数据(如关系型数据库、面向对象数据库中的数据)和完全无结构的数据(如声音、图像文件等)之间的数据,HTML文档就属于半结构化数据。它一般是自描述的,数据的结构和内容混在一起,没有明显的区分。

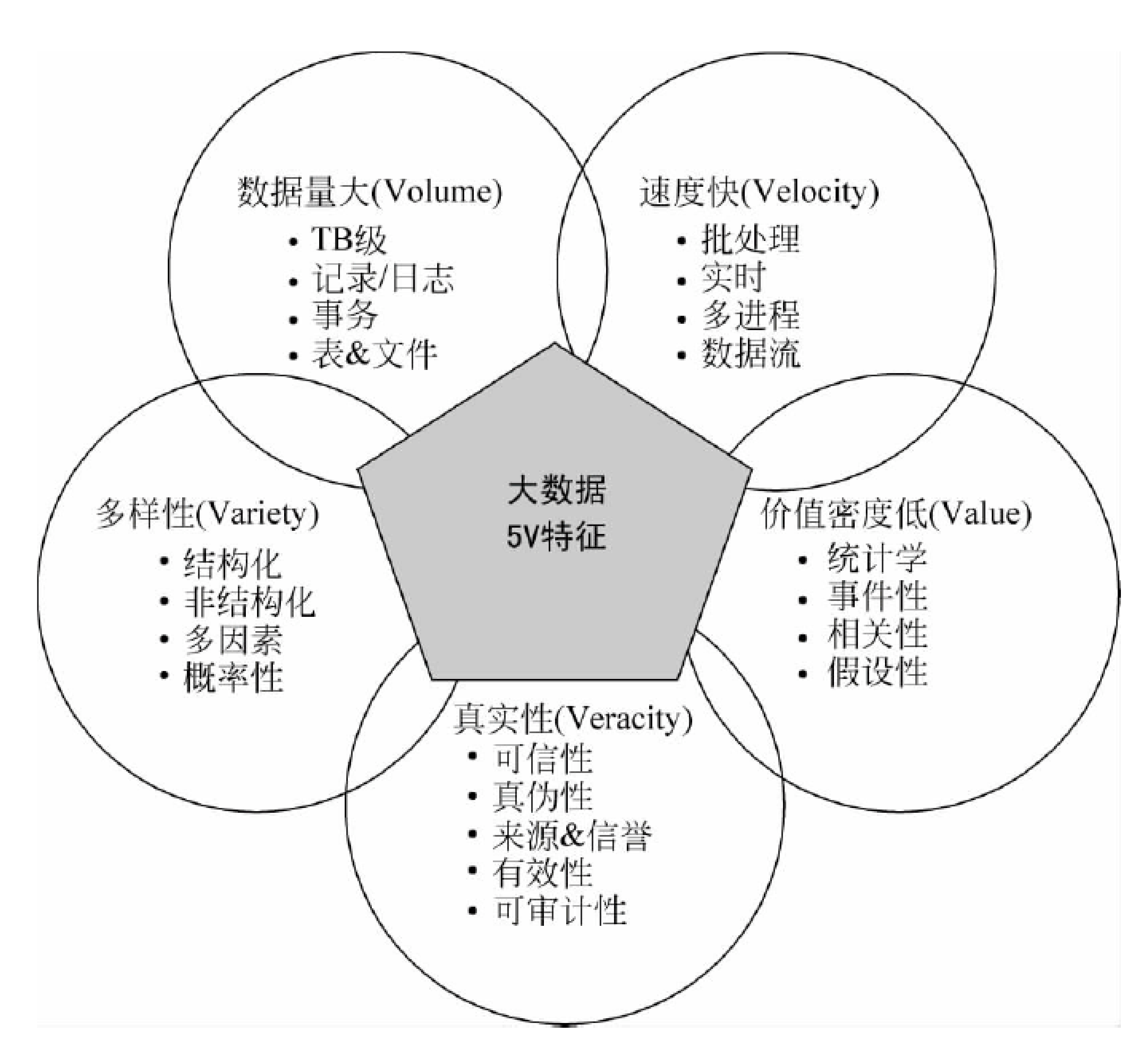


图 1-1 大数据 5V 特征

表 1-1 对大数据数据结构类型从多个角度进行了区分。

表 1-1 大数据数据结构类型区分

区别	类型		
	结构化数据	非结构化数据	半结构化数据
描述	包括预定义的数据类型、格式和结构的数据	没有固定结构的数据，通常保存为不同类型的文件	具有可识别的模式并可以解析的文本数据文件
数据实例	事务性数据和联机分析处理	文本、办公文档 PDF、图像、声音、视频等	自描述和具有定义模式的 XML 数据文件
数据模型	二维表(关系型)		树、图
访问	交互式和批处理	批处理	
数据大小	GB	PB	
结构	静态模式	动态模式	
模式	先有模式，再有数据	先有数据，再有模式	先有数据，再有模式
数据库	Oracle、Sybase、SQL Server、DB2、Informix 等	iBase，Hadoop，MapReduce，Hive	Hadoop，storm

随着网络技术的发展，特别是 Internet 和 Intranet 技术的飞快发展，使得非结构化数据的数量日趋增大。这时，主要用于管理结构化数据的关系数据库的局限性暴露得越来越明显。完全基于 Internet 应用的非结构化数据库将成为继层次数据库、网状数据库和关系数据库之后的又一重点、热点技术。因而，数据库技术相应地进入了“后关系数据库时代”，发展进入基于网络应用的非结构化数据库时代。

1.2.4 数据、信息、知识与智能的关系

数据、信息、知识是有相互关系又有区别的三个概念，正确理解它们之间的含义对于深

入理解大数据意义和价值具有重要作用。图 1-2 给出了数据、信息和知识的关系和区别。

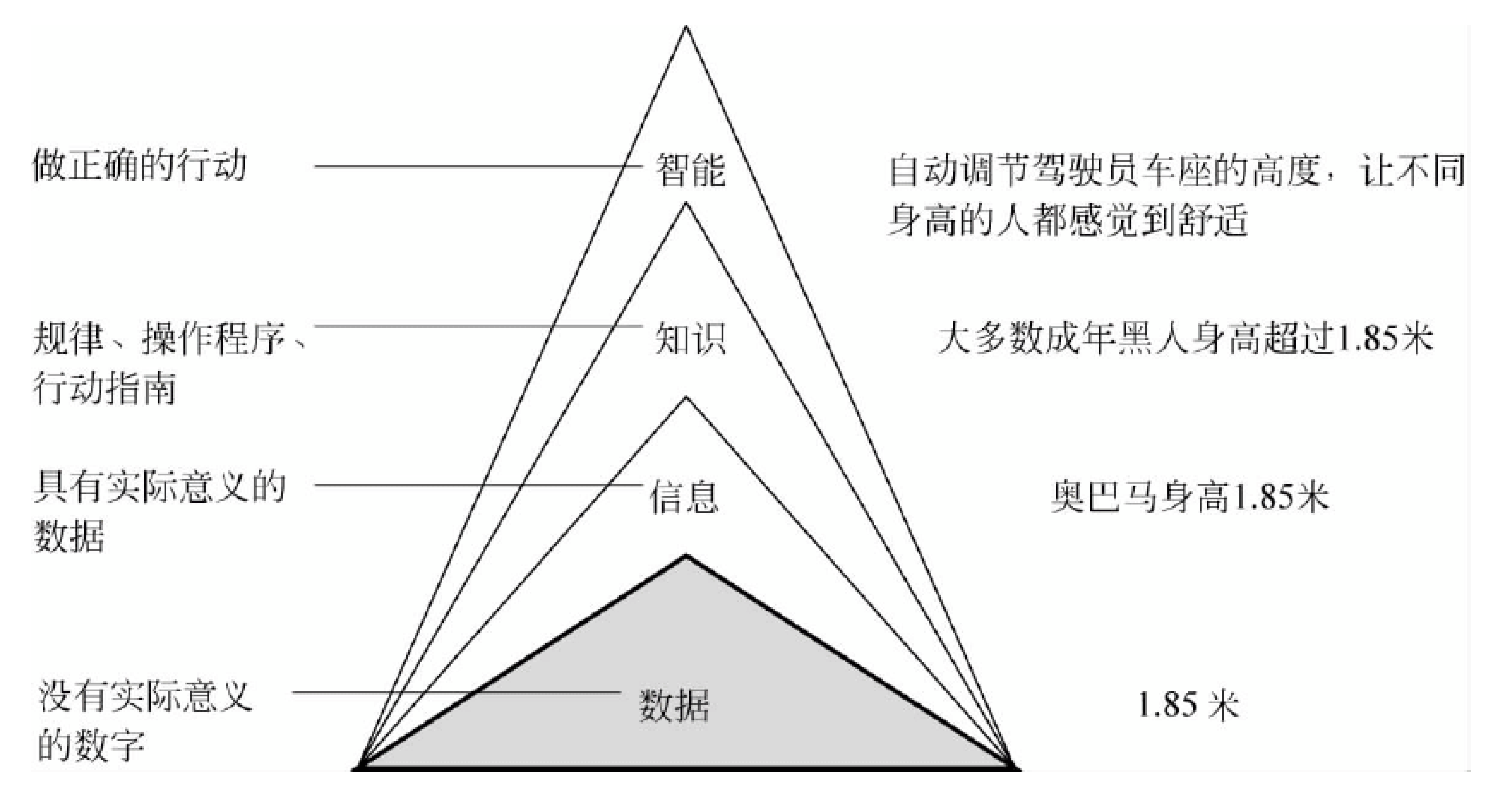


图 1-2　数据、信息、知识和智能的关系

数据是没有意义的数字，如 1.85 米，当人们看见 1.85 米时并不知道它是表示一个人的身高，还是游泳池的水深，必须将它与它所处的实际环境或场景相联系才能够准确地理解它。数据是最基础的元素，它反映了客观世界最基本的事实和运行状态。早期（1980 年之前）的信息化应用主要是做数据处理，计算机对数据进行处理，得到各种数字结果，然后再人为对这些结果进行解释，如绘制成特定形式的图或统计表。

信息是具有实际意义的数据，如“奥巴马身高 1.85 米”“小王喜欢音乐”。信息是通过对数据加工处理后得到的，信息是融入了人类数据处理和表现智慧后的数据呈现方式。目前，信息化应用主要是信息处理应用阶段，人们在计算机或者各种移动终端上得到的主要是信息，依据这些信息进行生产经营和业务决策。信息发挥的作用是提供竞争情报、企业经营状态、市场反馈消息、产品成本构成等资讯，帮助管理人员减少决策过程中的不确定性，但是如何理解这些信息，这些信息反映了客户世界的什么规律，如果做出相应的行动完全由人来决定，所做的决策和行动的正确性以及效果则完全因人而异。同样的信息，由于决策者的经验和认识不同，可能会做出截然不同的决策，甚至是背离事物运行方向的错误决策。基于数据和信息处理的信息技术应用属于信息化应用的初级阶段。

知识则是在信息的基础上，总结了人类实践敬仰后得到的对客观世界运行规律、操作程序和最佳行动策略的认识，如“大部分成年黑人的身高超过 1.85 米”就是一种知识。知识是主体获得的与客观事物存在及变化内在规律有关的系统化、组织化的信息。维基百科（http://zh.wikipedia.org/wiki/%E7%9F%A5%E8%AF%86）中对知识给出的定义是：知识是对某个主题确信的认识，并且这些认识拥有潜在的能力为特定目的而使用。知识是结构化的经验、价值、相关信息和专家洞察力的融合，提供了评价和产生新的经验和信息框架。维基百科的定义强调了知识的三个重要特性：第一，知识是确信的认知，是指知识是经过大量实践检验后形成的共识；第二，知识可以使用，知识的使用价值是知识的最大作用；第三，知识可以用来评价和产生新的知识，这就使得知识具有了生产要素的特性。随着信息技术的飞速发展，我们正在加速进入知识经济时代，在知识经济中知识成为重要的生产要素，所谓知识经济就是建立在知识的生产、分配和使用（消费）之上的经济。今天得到全球广

泛重视的大数据应用可以看作是知识应用的一个方面，即利用大数据挖掘技术，发现企业经营管理和市场运作中隐藏的规律，并用它来指导实践，获得市场竞争优势。企业不能满足当前的信息处理应用阶段的信息化应用现状，需要不断提升信息化应用的深度，向即将成为主流的知识应用阶段迈进，基于知识的信息技术应用是信息化应用的终极阶段。

智慧化是信息化应用的高级阶段。智慧是以知识和智能为基础，其中知识是一切智能行为的基础，而智力是获取知识并运用知识求解问题的能力，是头脑中思维活动的具体体现。智能是指个体对客观事物进行合理分析和判断，并灵活自适应地对变化的环境进行响应的一种能力。智能包括环境感知、逻辑推理、策略规划、行动和自学习 5 种能力，这 5 种能力是判断一个对象或系统是否具有智能的主要特征。这 5 种能力结合以后就可以形成若干种智能对象或系统，如智能机器人、智能汽车、智能调度与控制系统、智能工厂、智能停车场、智能电网等。下面以智能汽车为例对这 5 种能力进行介绍。

(1) 环境感知能力：具有对环境的基本模型建立功能，并能够感知到环境中的变化，如智能汽车可以感知到道路上的障碍物的交通信号灯的信息。

(2) 逻辑推理能力：运用所拥有的知识，对感知到的环境变化进行逻辑推理和判断，识别出对系统运行带来的影响，以决定是否需要采取必要行动，如智能汽车识别出信号灯是红色的，就需要停车，等信号灯变绿色后再启动汽车。

(3) 策略规划：在逻辑推理得出需要采用行动的情况下，策略规划功能负责制定一个最佳行动策略，如智能汽车识别出道路上的障碍物比较大，需要避让，策略规划功能根据当前的车速、邻近车道上是否有靠近的其他汽车、道路是否湿滑等情况，做出汽车减速和向左(向右)绕行路障的决策。

(4) 行动能力：按照策略规划功能给出的决策，执行系统进行行动操作，如智能汽车的油门和方向控制系统按照策略规划功能给出的策略控制汽车的行进速度方向。

(5) 自学习能力：每次执行行动完成后，对执行的结果进行评估，并总结经验，将成功的结果作为知识进行积累，对失败的结果作为反面案例知识也进行积累，通过学习和知识积累，系统不断进化，逐步对环境变化的响应速度和准确度越来越高。

《现代汉语词典》对智能的定义是"智慧和能力"，对智慧的定义是"辨析判断、发明创造的能力"。对智慧的另外一种定义是"对事物能迅速、灵活、正确地理解和处理的能力"。依据智慧的内容和所起作用的不同，可以把智慧分为三类：创造智慧、发现智慧和整合智慧。创造智慧，是指人们可以从无到有地创造和发明新东西的能力。发现智慧是指人们发掘已经存在但尚未被认知的事物或其本质、规律的能力。整合智慧是指人们运用现有的规则和知识来调整、梳理、矫正、改变已经存在的东西的能力。

帕梅拉·麦考达克(Pamela McCorduck)在她的著名的人工智能历史研究《机器思维》(*Machine Who Think*,1979)中曾经指出：在复杂的机械装置与智能之间存在着长期的联系。从几世纪前出现的神话般的复杂巨钟和机械自动机开始，人们已对机器操作的复杂性与自身的智能活动进行直接联系。

著名的英国科学家图灵被称为人工智能之父，图灵不仅创造了一个简单的通用的非数字计算模型，而且直接证明了计算机可能以某种被理解为智能的方法工作。1950 年，图灵发表了题为"计算机能思考吗?"的论文，给人工智能下了一个定义，而且论证了人工智能的可能性。定义智慧时，如果一台机器能够通过称为图灵实验的实验，那它就是智慧的。图灵

实验的本质就是让人在不看外形的情况下不能区别是机器的行为还是人的行为时，这个机器就是智慧的。

智能和智慧的主要区别，主要体现在以下几个方面。

（1）智慧更多地用于形容人，智能更多地用于形容物件或系统。

（2）智慧更多的是反映人类精神层面的活动过程，包括感知、综合、推理、判断、决策、学习等各种智力活动，它主要反映了人类拥有知识的丰富程度和认识事物本质的能力。

1.3　大数据发展史

早在 1970 年哈佛大学关于资源三角形的论述中，将材料、能源、信息看成是推动社会发展的三种基本资源。回顾过去的半个世纪，可以看到 IT 产业已经经历过几轮技术革命浪潮，每个阶段的浪潮都是由新兴的 IT 供应商主导，并极大地推动了信息技术和产业的发展。21 世纪是人类走向信息社会的世纪，是网络的时代，是超高速信息公路建设取得实质性进展并进入应用的年代。当前计算机正朝着巨型化、微型化、智能化、网络化等方向发展，计算机本身的性能越来越好，应用范围也越来越广，从而使计算机成为工作、学习和生活中必不可少的工具。数据来源于一切客观存在，包括宏观到微观的物理世界，各种生物体、人类社会活动、感知、认识和思维的结果。随着信息技术的发展，当通常所说的数据是指经过数字化转换后的信息，是可以被量化、分析和再利用的信息，包含数值、文字、符号、音频、视频等不同形态。对数据的分析如交通规划、宏观经济分析、电力系统规划、气象预测、高能物理、航天航空、基因工程等大规模数据分析和计算早已在人类生产和生活中发挥着关键的作用。

1.3.1　数据管理发展历程

随着计算机的发展，数据管理经历了几个重要的阶段。数据库技术从诞生到现在，在不到半个世纪的时间里，形成了坚实的理论基础、成熟的商业产品和广泛的应用领域，吸引越来越多的研究者加入。数据库的诞生和发展给计算机信息管理带来了一场巨大的革命。30 年间数据库领域获得了三次计算机图灵奖（C. W. Bachman，E. F. Codd，J. Gray），更加充分地说明了数据库是一个充满活力和创新精神的领域。下面就让我们沿着历史的轨迹，追溯一下数据库的发展历程，如图 1-3 数据管理技术发展历程所示。

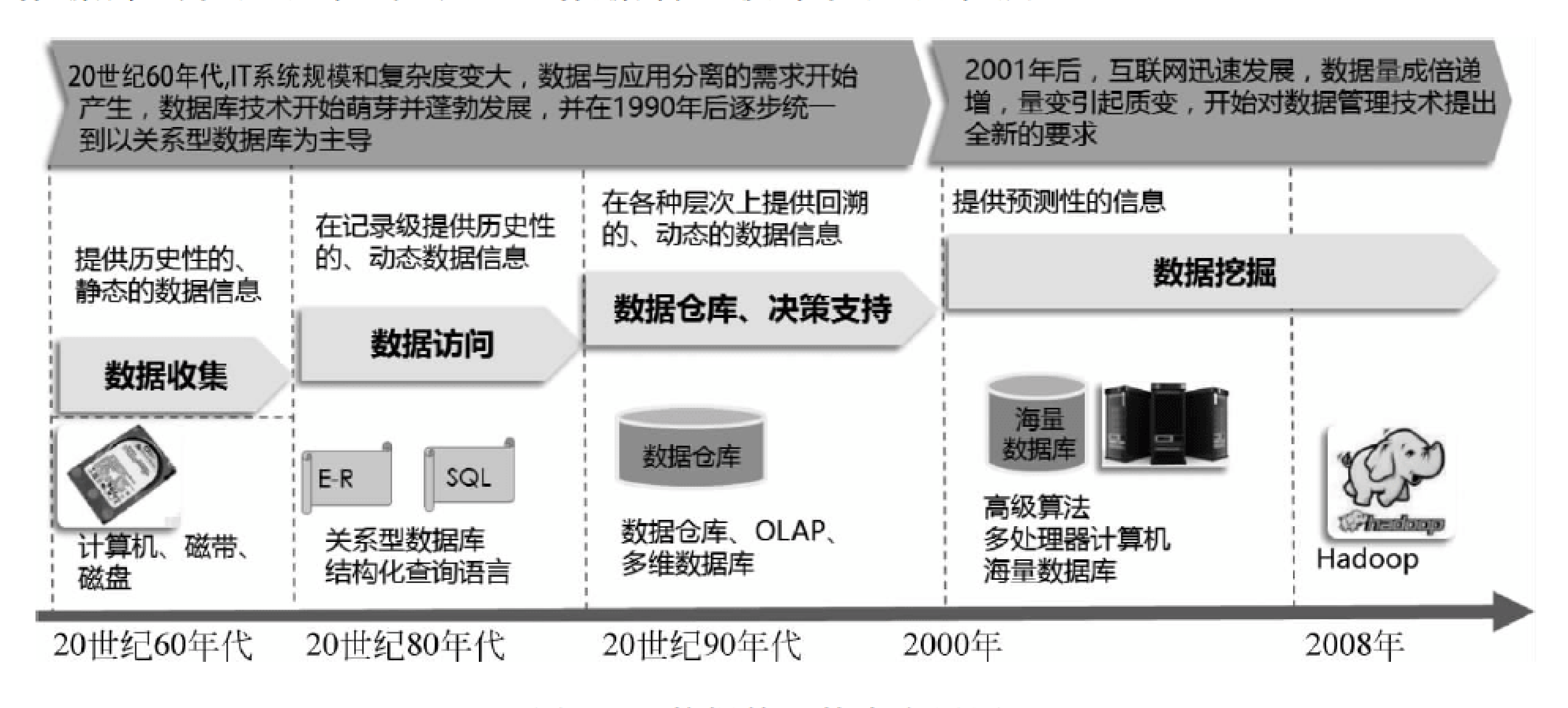


图 1-3　数据管理技术发展历程

1. 数据管理的诞生

20 世纪 60 年代，IT 系统规模和复杂度变大，数据与应用分离的需求开始产生，那时的数据管理非常简单，主要通过大量的分类、比较和表格绘制的机器运行数百万穿孔卡片来进行数据的处理，其运行结果在纸上打印出来或者制成新的穿孔卡片。而数据管理就是对所有这些穿孔卡片进行物理的储存和处理。

1951 年：Univac 系统使用磁带和穿孔卡片作为数据存储。1951 年，雷明顿兰德公司(Remington Rand Inc.)的一种叫做 Univac Ⅰ 的计算机推出了一种一秒钟可以输入数百条记录的磁带驱动器，从而引发了数据管理的革命。1956 年，IBM 生产出第一个磁盘驱动器—— the Model 305 RAMAC。此驱动器有 50 个盘片，每个盘片直径是二英尺，可以存储 5MB 的数据。使用磁盘最大的好处是可以随机地存取数据，而穿孔卡片和磁带只能顺序存取数据。

1961 年：通用电气(GE)公司的 Charles Bachman 开发了第一个数据库管理系统——IDS(Integrated Data Store)。1961 年，通用电气公司(General Electric Co.)的 Charles Bachman 成功地开发出世界上第一个网状 DBMS，也是第一个数据库管理系统——集成数据存储(Integrated Data Store IDS)，奠定了网状数据库的基础，并在当时得到了广泛的发行和应用。IDS 具有数据模式和日志的特征。但它只能在 GE 主机上运行，并且数据库只有一个文件，数据库所有的表必须通过手工编码来生成。之后，通用电气公司的一个客户(BF Goodrich Chemical 公司)最终不得不重写了整个系统，并将重写后的系统命名为集成数据管理系统(IDMS)。层次型 DBMS 是紧随网络型数据库而出现的。最著名最典型的层次数据库系统是 IBM 公司在 1968 年开发的 IMS(Information Management System)，一种适合其主机的层次数据库。这是 IBM 公司研制的最早的大型数据库系统程序产品。

2. 关系数据库的产生

由于计算机开始广泛地应用于数据管理，对数据的共享提出了越来越高的要求。传统的文件系统已经不能满足人们的需要，能够统一管理和共享数据的数据库管理系统(DBMS)应运而生。数据模型是数据库系统的核心和基础，各种 DBMS 软件都是基于某种数据模型的，所以通常也按照数据模型的特点将传统数据库系统分成网状数据库、层次数据库和关系数据库三类。网状数据库和层次数据库已经很好地解决了数据的集中和共享问题，但是在数据独立性和抽象级别上仍有很大欠缺。用户在对这两种数据库进行存取时，仍然需要明确数据的存储结构，指出存取路径。

1969 年：IBM 的研究员 Edgar F. Codd 博士发明了关系数据库。次年在刊物 *Communication of the ACM* 上发表了一篇名为 *A Relational Model of Data for Large Shared Data Banks* 的论文，提出了关系模型的概念，奠定了关系模型的理论基础。尽管在 1968 年 Childs 已经提出了面向集合的模型，然而这篇论文被普遍认为是数据库系统历史上具有划时代意义的里程碑。之后又陆续发表多篇文章，论述了范式理论和衡量关系系统的 12 条标准，用数学理论奠定了关系数据库的基础。

1974 年：IBM 的 Ray Boyce 和 Don Chamberlin 将 Codd 关系数据库的 12 条准则的数学定义以简单的关键字语法表现出来，里程碑式地提出了 SQL(Structured Query Language)。SQL 的功能包括查询、操纵、定义和控制，是一个综合的、通用的关系数据库语

言，同时又是一种高度非过程化的语言，只要求用户指出做什么而不需要指出怎么做。SQL 集成实现了数据库生命周期中的全部操作。SQL 提供了与关系数据库进行交互的方法，它可以与标准的编程语言一起工作。20 世纪 70 年代中期，关系理论通过 SQL 在商业数据库 Oracle 和 DB2 中使用。

1976 年：霍尼韦尔公司(Honeywell)公司推出了 Multics Relational Data Store——第一个商用关系数据库系统。关系型数据库系统以关系代数为坚实的理论基础，经过几十年的发展和实际应用，技术越来越成熟和完善。其代表产品有 Oracle、IBM 公司的 DB2、微软公司的 MS SQL Server 以及 Informix、ADABASD 等。

1979 年：Oracle 公司引入了第一个商用 SQL 关系数据库管理系统。

1983 年：IBM 推出了 DB2 数据库产品。

3. 数据仓库的形成

1985 年：为 Procter & Gamble 系统设计的第一个商务智能系统由 Metaphor 计算机系统有限公司为 Procter & Gamble 公司开发出来，主要是用来连接销售信息和零售的扫描仪数据。同年，Pilot 软件公司开始出售第一个商用客户/服务器执行信息系统——Command Center。

1991 年：W. H. Bill Inmon 发表了“构建数据仓库”。1988 年，IBM 公司的研究者 Barry Devlin 和 Paul Murphy 发明了一个新的术语——信息仓库，之后，IT 的厂商开始构建实验性的数据仓库。1991 年，W. H. Bill Inmon 出版了一本关于如何构建数据仓库的书，使得数据仓库真正开始应用。

4. 数据挖掘诞生

1997 年年底在加拿大温哥华举行的第五次亚太经合组织非正式首脑会议(APEC)上美国总统克林顿提出敦促各国共同促进电子商务发展的议案，引起了全球首脑的关注，IBM、HP 和 Sun 等国际著名的信息技术厂商宣布 1998 年为电子商务年。

随着互联网快速发展和数据库技术应用的不断深化，数据的积累不断膨胀，导致简单的查询和统计已经无法满足企业的商业需求，急需一些革命性的技术去挖掘数据背后的信息。同时，这期间计算机领域的人工智能(Artificial Intelligence)也取得了巨大进展，进入了机器学习的阶段。因此，人们将两者结合起来，用数据库管理系统存储数据，用计算机分析数据，并且尝试挖掘数据背后的信息。这两者的结合促生了一门新的学科，即数据库中的知识发现(Knowledge Discovery in Databases，KDD)。

1989 年 8 月召开的第 11 届国际人工智能联合会议的专题讨论会上首次出现了知识发现(KDD)这个术语，而数据挖掘(Data Mining)则是知识发现(KDD)的核心部分，它指的是从数据集合中自动抽取隐藏在数据中的那些有用信息的非平凡过程，这些信息的表现形式为：规则、概念、规律及模式等。进入 21 世纪，数据挖掘已经成为一门比较成熟的交叉学科，并且数据挖掘技术也伴随着信息技术的发展日益成熟起来。

数据挖掘融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的理论和技术，是 21 世纪初期对人类产生重大影响的十大新兴技术之一。

5. Hadoop 生态系统诞生

2005 年，Hadoop 最初只是雅虎公司用来解决网页搜索问题的一个项目，后来因其技术的高效性，被 Apache Software Foundation 公司引入并成为开源应用。2008 年，Hadoop 推出开源 1.0 发行版，其本身不是一个产品，而是由多个软件产品组成的一个生态系统，这些软件产品共同实现全面功能和灵活的大数据分析。从技术上看，Hadoop 由两项关键服务构成：采用 Hadoop 分布式文件系统(HDFS)的可靠数据存储服务，以及利用一种叫做 MapReduce 技术的高性能并行数据处理服务。这两项服务的共同目标是，提供一个使对结构化和复杂数据的快速、可靠分析变为现实的基础。

2008 年 6 月，思科发布了一份报告，题为“思科视觉网络指数——预测与方法，2007—2012”，作为“持续跟踪和预测视觉网络应用影响的行动”的一部分。这份报告预言，“从现在到 2012 年，IP 流量将每两年翻一番”，2012 年 IP 流量将达到 0.5ZB。这份预测比较准确，正如思科最近一份报告(2012 年 5 月 30 日)中指出的，2012 年 IP 流量刚刚超过 0.5ZB，“在过去的 5 年中增长了 8 倍”。

2008 年年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟(Computing Community Consortium)，发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不仅局限于数据处理的机器，并提出：大数据真正重要的是新用途和新见解，而非数据本身。此组织可以说是最早提出大数据概念的机构。

1.3.2 大数据的演变及回顾

2009 年，印度政府建立了用于身份识别管理的生物识别数据库，联合国全球脉冲项目已研究了对如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病暴发之类的问题。

2009 年年中，美国政府通过启动 Data.gov 网站的方式进一步开放了数据的大门，这个网站向公众提供各种各样的政府数据。该网站的超过 4.45 万的数据集被用于保证一些网站和智能手机应用程序来跟踪从航班到产品召回再到特定区域内失业率的信息，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

2009 年 12 月，罗杰·E. 博恩和詹姆斯·E. 少特发表了《信息知多少？2009 年美国消费者报告》。研究发现，2008 年“美国人消费了约 1.3 万亿小时信息，几乎平均每天消费 12 小时。总计 3.6 泽字节(ZB)，10 845 万亿单词，相当于平均每人每天消费 100 500 单词及 34GB 信息。”博恩、少特和沙坦亚·巴鲁在 2011 年 1 月发表了《信息知多少？2010 年企业服务器信息报告》，继续上述研究。在文中他们估计，2008 年“世界上的服务器处理了 9.57ZB 信息，几乎是 9.57×10^{22} 字节信息，或者是 10 万亿 GB。也就是平均每天每个工作者产生 12GB 信息，或者每年每个工作者产生 3TB 信息。世界上所有的公司平均每年处理 63TB 信息”。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。库克尔在报告中提到：“世界上有着无法想象的巨量数字信息，并以极快的速度增长。从经济界到科学界，从政府部门到艺术领域，很多方面都已经感受到了这种巨量信息的影响。科学家和计算机工程师已经为这个现象创造了一个新词汇：‘大数

据’。”库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2011 年 2 月，IBM 的沃森超级计算机每秒可扫描并分析 4TB(约两亿页文字量)的数据量，并在美国著名智力竞赛电视节目《危险边缘》(Jeopardy)上击败两名人类选手而夺冠。后来纽约时报认为这一刻为一个“大数据计算的胜利”。

2011 年 5 月，全球知名咨询公司麦肯锡(McKinsey&Company)全球研究院(MGI)发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，从此大数据开始备受关注，这也是专业机构第一次全方面介绍和展望大数据。报告指出，大数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。报告还提到，“大数据”源于数据生产和收集的能力和速度的大幅提升——由于越来越多的人、设备和传感器通过数字网络连接起来，产生、传送、分享和访问数据的能力也得到彻底变革。

2011 年 12 月，工业和信息化部发布的物联网十二五规划中，把信息处理技术作为 4 项关键技术创新工程之一被提出来，其中包括海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。

2012 年 1 月份，瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响》(*Big Data*，*Big Impact*)宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。

2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。2012 年 3 月 22 日，奥巴马政府宣布两亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭，在次日的电话会议中，政府将数据定义为“未来的新石油”，大数据技术领域的竞争，事关国家安全和未来。并表示，国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用数据的能力；国家数字主权体现对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。

2012 年 4 月，美国软件公司 Splunk 于 19 日在纳斯达克成功上市，成为第一家上市的大数据处理公司。鉴于美国经济持续低迷、股市持续震荡的大背景，Splunk 首日的突出交易表现尤其令人们印象深刻，首日即暴涨了一倍多。Splunk 是一家领先的提供大数据监测和分析服务的软件提供商，成立于 2003 年。Splunk 成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快大数据布局。

2012 年 7 月，联合国在纽约发布了一份关于大数据政务的白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求：例如，通过对价格关注和更好服务的渴望，个人提供数据和众包信息，并对隐私和退出权力提出需求；公共部门出于改善服务，提升效益的目的，提供了诸如统计数据、设备信息、健康指标，及税务和消费信息等，并对隐私和退出权力提出需求；私人部门出于提升客户认知和预测趋势目的，提供汇总数据、消费和使用信息，并对敏感数据所有权和商业模式更加关注。白皮书还指出，人们如今可以使用的极大丰富的数据资源，包括旧数据和新数据，来对社会人口进行前所未有的实时分析。联合国还以爱尔兰和美国的社交网络活跃度增长可以作为失业率上升的早期征兆为例，表明政府如果能合理分析所掌握的数据资源，将能“与数俱进”，快速应变。

2012 年 12 月“世界经济论坛”发布《大数据、大影响》报告，阐述大数据在金融服务、健康、教育、农业、医疗等多个领域给世界经济社会发展带来的机会。

2013 年 5 月，麦肯锡全球研究所（McKinsey Global Institute）发布了一份名为《颠覆性技术：技术进步改变生活、商业和全球经济》的研究报告。报告确认的未来 12 种新兴技术，有望在 2025 年带来 14 万亿至 33 万亿美元的经济效益。令人惊讶的是，最为热门的大数据技术却未被列入其中。麦肯锡专门解释称，大数据已成为这些可能改变世界格局的 12 项技术中许多技术的基石，包括移动互联网、知识工作自动化、物联网、云计算、先进机器人、自动汽车、基因组学等都少不了大数据应用。

2014 年 4 月，世界经济论坛以“大数据的回报与风险”主题发布了《全球信息技术报告（第 13 版）》。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要。在接下来将对数据保密和网络管制等议题展开积极讨论。全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到大数据在推动经济发展、改善公共服务，增进人民福祉，乃至保障国家安全方面的重大意义。

2014 年 5 月，美国白宫发布了 2014 年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。

2015 年 8 月国务院发布《促进大数据发展行动纲要》，这为我国大数据发展进行了顶层设计和统筹部署，这是目前为止我国促进大数据发展的第一份权威性、系统性文件，从国家大数据发展战略全局的高度，提出了我国大数据发展的顶层设计，是指导我国未来大数据发展的纲领性文件。2015 年 10 月 31 日，十八届五中全会通过《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》，规划指出：实施网络强国战略，实施“互联网＋”行动计划，发展分享经济，实施国家大数据战略。

2015 年 11 月大数据产业“十三五”发展规划编制小组在京召开专题研讨会，着手编制《大数据产业“十三五”发展规划》，将其作为贯彻国家大数据战略、落实《促进大数据发展行动纲要》、加快建设数据强国的重要抓手。除制定《大数据产业“十三五”发展规划》外，工信部还将出台促进大数据产业发展的推进计划，促进规划、标准、技术、产业、安全、应用的协同发展。

2016 年 1 月国务院发布《关于组织实施促进大数据发展重大工程的通知》，提出加快落实《大数据纲要》，从破解制约大数据创新发展的突出矛盾和问题出发，重点推进数据资源开放共享，推动大数据基础设施统筹，打破数据资源壁垒，深化数据资源应用，积极培育新兴繁荣的产业发展新业态。同时通知提到重点支持大数据示范应用、重点支持大数据共享开放、重点支持基础设施统筹发展、重点支持数据要素流通。

2016 年 5 月 25 日，由国家发展改革委员会、贵州省政府共同主办的“国家级”大数据行业盛宴——贵阳数博会就要拉开序幕了。这个在全球科技界范围内都有影响力的盛会，2016 年恰逢身处人工智能、虚拟现实、机器学习等高新科技热潮风口，因而更加意义非凡。

1.4　大数据的度量和价值

1.4.1　大数据的度量

数据量的大小是用计算机存储容量的单位来计算的，基本的单位的是字节(Byte)，每一级按照千分位递进，如下所示。

1Byte(B)＝8bits	一个英文字母占用空间
1KiloByte(KB)＝1024B	相当于一则短篇故事的内容
1MegaByte(MB)＝1024KB	相当于一则短篇小说的文字内容
1GigaByte(GB)＝1024MB	相当于贝多芬第五乐章交响曲的乐谱内容
1TeraByte(TB)＝1024GB	相当于一家大型医院中所有的 X 光图片内容
1PetaByte(PB)＝1024TB	相当于 50%的全美学术研究图书馆藏书信息内容
1ExaByte(EB)＝1024PB	5EB 相当于至今全世界人类所讲过的话
1ZettaByte(ZB)＝1024EB	截至 2010 年，人类拥有的信息总量是 1.2ZB
1YottaByte(YB)＝1024ZB	
1BrontoByte(BB)＝1024YB	

1.4.2　大数据的价值

研究表明，数据的价值会随着时间的流逝而降低。简单地说，数据的价值与时间是成反比的。因此，数据处理速度越快，数据价值越能够更好地获取。大数据的价值也与它所传播与共享的范围有关，使用大数据的用户越多，范围越广，信息的价值就越大。大数据价值的充分发挥，依赖于大数据的分析和挖掘技术，更好的分析工具和算法能够获得更为准确的信息，也更能发挥其价值。总之，大数据的价值，可以用如下公式来简单定义：

$$\text{大数据价值 } V=\frac{\text{大数据处理以及分析算法和工具}\int(\text{大数据量 } v_1\text{，大数据种类 } v_2\text{，高速流动 } v_3)}{\text{大数据存在时间}}\times\text{大数据用户数}$$

因此，大数据处理和分析技术对于挖掘大数据价值的作用十分关键。

据资料显示，近年来，甲骨文、IBM、微软、SAP、惠普等公司已经在数据管理和分析领域投入超出 150 亿美元。

大数据在以下 5 个方面创造价值。

(1) 先见之明——通过已经发生的、正在发生的事件或实验结果发现或预测需求，洞察变化倾向。

(2) 英明决策——自动算法代替/支持人类的决策。

(3) 一目了然——发现数据之间的关系。

(4) 有的放矢——细分人群，定制行动。

(5) 推陈出新——创新的商业模式、产品和服务。

(摘自麦肯锡《大数据：创新、竞争和提高生产率的下一个新领域》)

1. 改变经济社会管理方式

大数据作为一种重要的战略资产，已经不同程度地渗透到每个行业领域和部门，其深度

应用不仅有助于企业经营活动，还有利于推动国民经济发展。在宏观层面，大数据使经济决策部门可以更敏锐地把握经济走向，制定并实施科学的经济政策。在微观层面，大数据可以提高企业经营决策水平和效率，推动创新，给企业、行业领域带来价值。大数据技术作为一种重要的信息技术，对于提高安全保障能力、应急能力、优化公共事业服务、提高社会管理水平的作用正在日益凸显；在国防、反恐、安全等领域，应用大数据技术能够对来自于多种渠道的信息快速进行自动分类、整理、分析和反馈，有效解决情报、监视和侦察系统不足等问题，提高国家安全保障能力。

除此之外，大数据还将推动社会各个主体共同参与社会治理。网络社会是一个复杂、开放的巨系统，这个巨系统打破了传统组织的层级化结构，呈现出扁平化特征。个体的身份经历了从单位人、社会人到网络人的转变过程。政府、企业、社会组织、公民等各种主体都以更加平等的身份参与到网络社会的互动和合作之中，这对促进城市转型升级和提高可持续发展能力、提升社会治理能力、实现推进社会治理机制创新、促进社会治理实现管理精细化、服务智慧化、决策科学化、品质高端化等具有重要作用。

2. 促进行业融合发展

网络环境、移动终端随影而行，网上购物、社交网站、电子邮件、微信不可或缺，社会主体的日常活动在虚拟的环境下得到承载和体现。正如工业化时代商品和交易的快速流通催生大规模制造业发展，信息的大量、快速流通将伴随着行业的融合发展，使经济形态发生大范围变化。

大数据应用的关键在于分享，各行业已逐渐意识到单一的数据是没法发挥最大效能的，行业或部门之间相互交换数据已经成为一种发展趋势。虚拟环境下，遵循类似摩尔定律原则增长的海量数据，在技术和业务的促进下，使跨领域、跨系统、跨地域的数据共享成为可能，大数据支持着机构业务决策和管理决策的精准性、科学性以及社会整体层面的业务协同效率提高。

3. 推动产业转型升级

信息消费作为一种以信息产品和服务为消费对象的活动，覆盖多种服务形态、多种信息产品和多种服务模式。当围绕数据的业务在数据规模、类型和变化速度达到一定程度时，大数据对于产业发展的影响随之显现。

在面对多维度、爆发式增长的海量数据时，ICT 产业面临着有效存储、实时分析、高性能计算等挑战，这将对软件产业、芯片以及存储产业产生重要影响，推动一体化数据存储处理服务器、内存计算等产品的升级创新。对数据快速处理和分析的需求，将推动商业智能、数据挖掘等软件在企业级的信息系统中得到融合应用，成为业务创新的重要手段。

同时，“互联网＋”战略使大数据在促进网络通信技术与传统产业密切融合方面的作用更加凸显，对于传统产业的转型发展，创造更多价值影响重大。未来，大数据发展将不仅催生软硬件及服务等市场产生大量价值，也将对有关的传统行业转型升级产生重要影响。

4. 助力智慧城市建设

信息资源开发利用水平，在某种程度上代表着信息时代下社会的整体发展水平和运转效率。大数据与智慧城市是信息化建设的内容与平台，两者互为推动力量。智慧城市是大数据的源头，大数据是智慧城市的内核。

针对政府,大数据为政府管理提供强大的决策支持。在城市规划方面,通过对城市地理、气象等自然信息和经济、社会、文化、人口等人文社会信息的挖掘,可以为城市规划提供强大的决策支持,强化城市管理服务的科学性和前瞻性;在交通管理方面,通过对道路交通信息的实时挖掘,能够有效缓解交通拥堵,并快速响应突发状况,为城市交通的良性运转提供科学的决策依据;在舆情监控方面,通过网络关键词搜索及语义智能分析,能提高舆情分析的及时性、全面性,全面掌握社情民意,提高公共服务能力,应对网络突发的公共事件,打击违法犯罪;在安防领域,通过大数据的挖掘,可以及时发现人为或自然灾害、恐怖事件,提高应急处理能力和安全防范能力。

针对民生,大数据将提高城市居民的生活品质。与民生密切相关的智慧应用包括智慧交通、智慧医疗、智慧家居、智慧安防等,这些智慧化的应用将极大地拓展民众生活空间,引领大数据时代智慧人生的到来。大数据是未来人们享受智慧生活的基础,将改变传统"简单平面"的生活常态,通过大数据的应用服务将使信息变得更加泛在,使生活变得多维和立体。

5. 创新商业模式

大数据时代,产业发展模式和格局正在发生深刻变革。围绕着数据价值的行业创新发展将悄然影响各行各业的主营业态。而随之带来的,则是大数据产业下的创新商业模式。

一方面围绕数据产品价值链而产生诸如数据租售模式、信息租售模式、知识租售模式等。数据租售旨在为客户提供原始数据的租售;信息租售旨在向客户租售某种主题的相关数据集,是对原始数据进行整合、提炼、萃取,使数据形成价值密度更高的信息;知识租售旨在为客户提供一体化的业务问题解决方案,是将原始数据或信息与行业知识利用相结合,通过行业专家深入介入客户业务流程,提供业务问题解决方案。

另一方面,通过对大数据的分析处理,企业现有的商业模式、业务流程、组织架构、生产体系、营销体系也将发生变革。以数据为中心,挖掘客户潜在需求,不仅能够提升企业运作的效率,更可以藉由数据重新思考商业社会的需求与自身业务模式的转型,快速重构新的价值链,建立新的行业领导能力,提升企业影响力。

6. 改变科学研究的方法论

大数据技术的兴起对传统的科学方法论带来了挑战和革命。随着计算技术和网络技术的发展,采集、存储、传输和处理数据都已经成了容易实现的事情。面对复杂对象,我们没有必要再做过多的还原和精简,而是可以通过大量数据甚至是海量数据来全面、完整地刻画对象,通过处理海量数据来找到研究对象的规律或本质。当数据处理技术已经发生翻天覆地的变化时,在大数据时代我们需要的是所有数据,即"样本=总体",相比依赖于小数据和精确性的时代,大数据因为更强调数据的完整性和混杂性,突出事物的关联性,为我们解决问题提供新的视角,帮助我们进一步接近事实的真相。

1.5　大数据生态系统

1.5.1　大数据生态系统全貌

图 1-4 是 Big Data Group 所描绘的大数据云图,从图中可以看出,围绕大数据已经逐渐演化发展成为十分繁荣的生态系统,里面包括提供硬件、操作系统软件、数据库软件、应用软

件、云平台软件、数据分析、咨询服务等各种类型业务的公司，这些公司在大数据基础设施层和应用层分别提供不同类型的服务，同类型服务之间的相互竞争，不同类型的服务之间相互协作，共同形成一个以大数据为核心的服务协同生态系统。

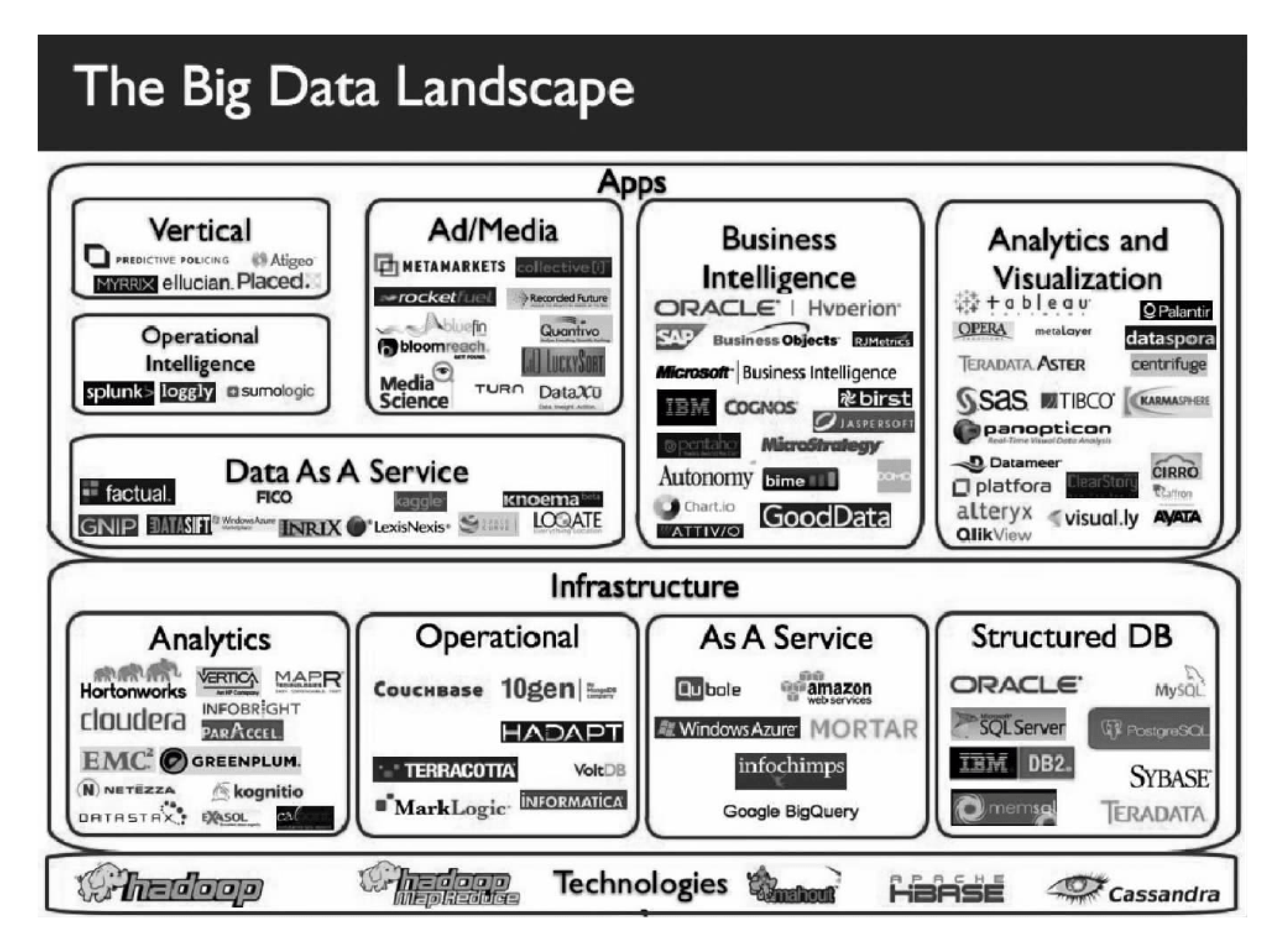


图 1-4　Big Data Group 所描绘的大数据云图

大数据生态系统的分类，底层是基础架构提供商和大数据平台提供商，这两类还是以传统的软硬件厂商为主。而上层是专业服务商和应用服务商。这两类企业都是直接面对最终用户的，但盈利模式有很大的区别。专业服务商的业务专业性比较强，通常是跨大数据领域的多个环节的；而应用服务商则大部分是在传统应用里嵌入大数据的概念或技术在业务模式上，专业服务商是偏重于运营，而应用服务商以项目型为主。

1.5.2　大数据生态系统框架

如图 1-5 所示为 Hadoop 大数据生态系统框架。

1. MapReduce 并行计算框架

MapReduce 并行计算框架是一个并行化程序执行系统。它提供了一个包含 Map 和 Reduce 两阶段的并行处理模型和过程，提供一个并行化编程模型和接口，让程序员可以方便快速地编写出大数据并行处理程序。MapReduce 以键值对数据输入方式来处理数据，并能自动完成数据的划分和调度管理。在程序执行时，MapReduce 并行计算框架将负责调度和分配计算资源，划分和输入输出数据，调度程序的执行，监控程序的执行状态，并负责程序执行时各计算节点的同步以及中间结果的收集整理。MapReduce 框架提供了一组完整的供程序员开发 MapReduce 应用程序的编程接口。

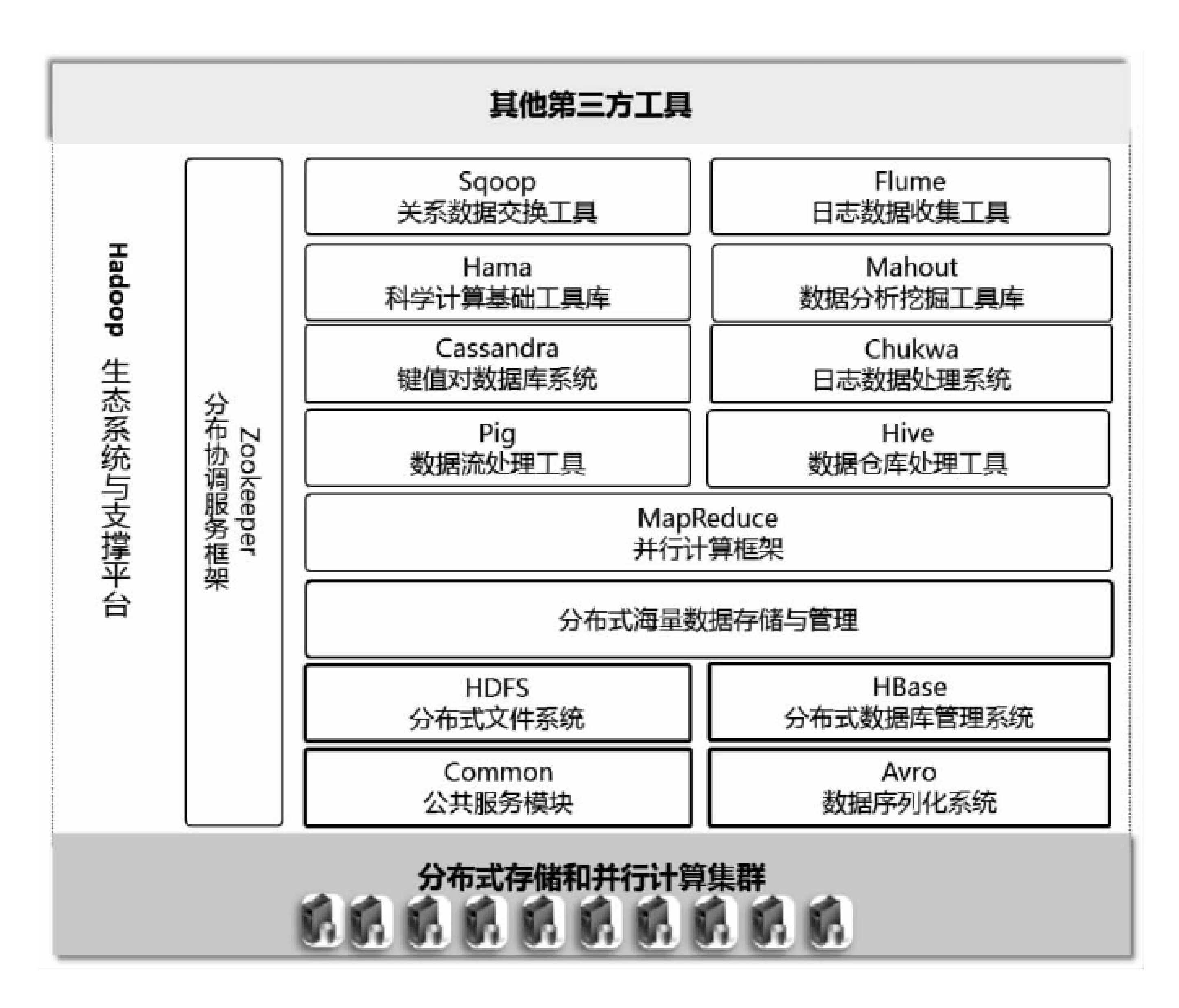


图 1-5　大数据生态系统框架

2. 分布式文件系统

HDFS(Hadoop Distributed File System，分布式文件系统)是一个类似于 Google GFS 的开源的分布式文件系统。它提供了一个可扩展、高可靠、高可用的大规模数据分布式存储管理系统，基于物理上分布在各个数据存储节点的本地 Linux 系统的文件系统，为上层应用程序提供了一个逻辑上成为整体的大规模数据存储文件系统。与 GFS 类似，HDFS 采用多副本(默认为三个副本)数据冗余存储机制，并提供了有效的数据出错检测和数据恢复机制，大大提高了数据存储的可靠性。

3. 分布式数据库管理系统

为了克服 HDFS 难以管理结构化/半结构化海量数据的缺点，Hadoop 提供了一个大规模分布式数据库管理和查询系统 HBase。HBase 是一个建立在 HDFS 之上的分布式数据库，它是一个分布式可扩展的 NoSQL 数据库，提供了对结构化、半结构化甚至非结构化大数据的实时读写和随机访问能力。HBase 提供了一个基于行、列和时间戳的三维数据管理模型，HBase 中每张表的记录数(行数)可以多达几十亿条甚至更多，每条记录可以拥有多达上百万的字段。

4. 公共服务模块

公共服务模块(Common)是一套为整个 Hadoop 系统提供底层支撑服务和常用工具的类库和 API 编程接口，这些底层服务包括 Hadoop 抽象文件系统 FileSystem、远程过程调用 RPC、系统配置工具 Configuration 以及序列化机制。在 0.20 及以前的版本中，Common 包含 HDFS、MapReduce 和其他公共的项目内容；从 0.21 版本开始，HDFS 和 MapReduce 被分离为独立的子项目，其余部分内容构成 Hadoop Common。

5. 数据序列化系统

数据序列化系统(Avro)是一个数据序列化系统，用于将数据结构或数据对象转换成便

于数据存储和网络传输的格式。Avro 提供了丰富的数据结构类型,快速可压缩的二进制数据格式,存储持久性数据的文件集,远程调用 RPC 和简单动态语言集成等功能。

6. 分布式协调服务框架

分布式协调服务框架(Zookeeper)是一个分布式协调服务框架,主要用于解决分布式环境中的一致性问题。Zookeeper 主要用于提供分布式应用中经常需要的系统可靠性维护、数据状态同步、统一命名服务、分布式应用配置项管理等功能。Zookeeper 可用来在分布式环境下维护系统运行管理中的一些数据量不大的重要状态数据,并提供监测数据状态变化的机制,以此配合其他 Hadoop 子系统(如 HBase、Hama 等)或者用户开发的应用系统,解决分布式环境下系统可靠性管理和数据状态维护等问题。

7. 分布式数据仓库处理工具

分布式数据仓库处理工具(Hive)是一个建立在 Hadoop 之上的数据仓库,用于管理存储于 HDFS 或 HBase 中的结构化/半结构化数据。它最早由 Facebook 开发并用于处理并分析大量的用户及日志数据,2008 年 Facebook 将其贡献给 Apache 成为 Hadoop 开源项目。为了便于熟悉 SQL 的传统数据库使用者使用 Hadoop 系统进行数据查询分析,Hive 允许直接用类似 SQL 的 HiveQL 查询语言作为编程接口编写数据查询分析程序,并提供数据仓库所需要的数据抽取转换、存储管理和查询分析功能,而 HiveQL 语句在底层实现时被转换为相应的 MapReduce 程序加以执行。

8. 数据流处理工具

数据流处理工具(Pig)是一个用来处理大规模数据集的平台,由 Yahoo! 贡献给 Apache 成为开源项目。它简化了使用 Hadoop 进行数据分析处理的难度,提供一个面向领域的高层抽象语言 Pig Latin,通过该语言,程序员可以将复杂的数据分析任务实现为 Pig 操作上的数据流脚本,这些脚本最终执行时将被系统自动转换为 MapReduce 任务链,在 Hadoop 上加以执行。Yahoo! 有大量的 MapReduce 作业是通过 Pig 实现的。

9. 键值对数据库系统

键值对数据库系统(Cassandra)是一套分布式的 K-V 型的数据库系统,最初由 Facebook 开发,用于存储邮箱等比较简单的格式化数据,后 Facebook 将 Cassandra 贡献出来成为 Hadoop 开源项目。Cassandra 以 Amazon 专有的完全分布式 Dynamo 为基础,结合了 Google BigTable 基于列族(Column Family)的数据模型,提供了一套高度可扩展、最终一致、分布式的结构化键值存储系统。它结合了 Dynamo 的分布技术和 Google 的 BigTable 数据模型,更好地满足了海量数据存储的需求。同时,Cassandra 变更垂直扩展为水平扩展,相比其他典型的键值数据存储模型,Cassandra 提供了更为丰富的功能。

10. 日志数据处理系统

日志数据处理系统(Chukwa)是一个由 Yahoo!贡献的开源的数据收集系统,主要用于日志的收集和数据的监控,并与 MapReduce 协同处理数据。Chukwa 是一个基于 Hadoop 的大规模集群监控系统,继承了 Hadoop 系统的可靠性,具有良好的适应性和扩展性。它使用 HDFS 来存储数据,使用 MapReduce 来处理数据,同时还提供灵活强大的辅助工具用以分析、显示、监视数据结果。

11. 科学计算基础工具库

科学计算基础工具库(Hama)是一个基于 BSP 并行计算模型(Bulk Synchronous Parallel,大同步并行模型)的计算框架,主要提供一套支撑框架和工具,支持大规模科学计算或者具有复杂数据关联性的图计算。Hama 类似 Google 公司开发的 Pregel,Google 利用 Pregel 来实现图遍历(BFS)、最短路径(SSSP)、PageRank 等计算。Hama 可以与 Hadoop 的 HDSF 进行完美的整合,利用 HDFS 对需要运行的任务和数据进行持久化存储。由于 BSP 在并行化计算模型上的灵活性,Hama 框架可在大规模科学计算和图计算方面得到较多应用,完成矩阵计算、排序计算、PageRank、BFS 等不同的大数据计算和处理任务。

12. 数据分析挖掘工具库

数据分析挖掘工具库(Mahout)来源于 Apache Lucene 子项目,其主要目标是创建并提供经典的机器学习和数据挖掘并行化算法类库,以便减轻需要使用这些算法进行数据分析挖掘的程序员的编程负担,不需要自己再去实现这些算法。Mahout 现在已经包含聚类、分类、推荐引擎、频繁项集挖掘等广泛使用的机器学习和数据挖掘算法。此外,它还提供了包含数据输入输出工具,以及与其他数据存储管理系统进行数据集成的工具和构架。

13. 关系数据交换工具

关系数据交换工具(Sqoop)是 SQL-to-Hadoop 的缩写,是一个在关系数据库与 Hadoop 平台间进行快速批量数据交换的工具。它可以将一个关系数据库中的数据批量导入 Hadoop 的 HDFS、HBase、Hive 中,也可以反过来将 Hadoop 平台中的数据导入关系数据库中。Sqoop 充分利用了 Hadoop MapReduce 的并行化优点,整个数据交换过程基于 MapReduce 实现并行化的快速处理。

14. 日志数据收集工具

日志数据收集工具(Flume)是由 Cloudera 开发维护的一个分布式、高可靠、高可用、适合复杂环境下大规模日志数据采集的系统。它将数据从产生、传输、处理、输出的过程抽象为数据流,并允许在数据源中定义数据发送方,从而支持收集基于各种不同传输协议的数据,并提供对日志数据进行简单的数据过滤、格式转换等处理能力。输出时,Flume 可支持将日志数据写往用户定制的输出目标。

1.6　大数据应用研究方向

当前“大数据”这一术语已经远远超越了当初的互联网或信息技术(IT)的技术范畴,变成了一个时代的标志。大数据时代的到来有其必然性,当计算和通信取得长足进步的时候,当传感器网络和互联网等信息采集平台日臻完善的时候,数据的存储管理和分析处理就自然成为关注的焦点。“大数据”概念的提出意味着信息技术领域的重点由“计算”转为“数据”。原来的“计算机科学”也正在潜移默化地向“数据科学”转化。大数据在科学研究(如地球科学、生命科学、高能物理研究等)和商业领域(如行为分析、趋势分析、行情预测、精准营销、商品推荐等)都有成功的应用。互联网已经成为人们生活生产中不可或缺的环境和平台,正因为大数据在互联网商业领域的巨大成功,使得这一概念已经被社会各个层面广泛认可,开始从线上走到线下,越来越多的人从企业管理、社会治理、科学研究等领域探讨大数据

的应用。这种来源于应用的关于大数据技术的爆发式需求，推动了数据科学的发展，因为其“应用驱动”的特点，工程实现和应用部署至关重要，“数据科学与工程”这一学科名称自然应运而生。因此，基于以上的认识，大数据可以细分为大数据管理、大数据技术、大数据科学和大数据工程几个重点方向，后面章节将围绕这几个方向进行详细阐述。根据这几个方向的应用相关性，本书将按照大数据管理与技术、大数据科学与工程两大类进行阐述。

1.6.1 大数据管理与技术

大数据的出现必将颠覆传统的数据管理方式，在数据来源、数据处理方式和数据思维等方面都会对其带来革命性的变化。数据思维要从以计算为中心转变到以数据处理为中心，这种方式需要我们从根本上转变思维。

传统数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的过程。其目的在于充分有效地发挥数据的作用。实现数据有效管理的关键是数据组织。随着计算机技术的发展，传统数据管理经历了人工管理、文件系统、数据库系统三个发展阶段。在数据库系统中所建立的数据结构，更充分地描述了数据间的内在联系，便于数据修改、更新与扩充，同时保证了数据的独立性、可靠、安全性与完整性，减少了数据冗余，故提高了数据共享程度及数据管理效率。

而大数据管理是指数据大小、形态超出典型数据管理系统采集、储存、管理和分析等能力的大规模数据集，而且这些数据之间存在着直接或间接的关联性，通过大数据技术可以从中挖掘出模式与知识。大数据技术是使大数据中所蕴含的价值得以挖掘和展现的一系列技术与方法，包括数据采集、预处理、存储、分析挖掘、可视化等。大数据应用，是对特定的大数据集、集成应用大数据系列技术与方法，获得有价值信息的过程。大数据技术的研究与突破，其最终目标就是从复杂的数据集中发现新的模式与知识，挖掘得到有价值的新信息。数据的量越来越大，种类越来越丰富，大数据时代需要新的数据管理手段。列式、MPP 的关系型数据仓库在改变着，NoSQL 的 CDBMS、GDBMS 也试图在改变着。关系型数据库是企业 IT 建设时代的数据管理基石，而在 Big Data 时代，也许需要一种新的，正在探索中的数据管理基石。目前，典型代表工具主要是 Hadoop 生态系统相关技术、Storm 等。

大数据管理可以更好地帮助人们对数据进行分类、归类；更好地优化资源，更好地识别和预测行为。在上述基础上，大数据也日益和分类、决策、预测等人们的行为相互渗透，以致人们自己也嵌入到大数据中，影响人们的行为。

1.6.2 大数据科学与工程

数据科学通常指基于计算机科学、统计学、信息系统等学科的理论和技术，研究数据的收集整理以及从海量数据中分析处理，获得有效知识并加以应用的新兴学科；数据工程是指利用工程的观点进行数据管理和分析以及开展系统的研发和应用。数据量的爆炸式增长不但改变了人们的生活方式、企业的运营模式，也改变了科学研究的基本范式。数据科学和工程可以作为支撑大数据研究与应用的交叉学科，其理论基础来自多个不同的学科领域，包括计算机科学、统计学、人工智能、信息系统、情报科学等。

与传统计算机和软件工程等学科相比，“数据科学与工程”更具备独特的学科基础和内涵。数据科学与工程学科的理论基础涉及统计分析、商务智能以及数据处理基础，具体包括

以下几个方面。

(1) 大数据表达理论方面：包括大数据的生命周期、演化与传播规律、数据科学与社会学、经济学等之间的互动机制以及大数据的结构与效能的规律性。

(2) 在大数据计算理论方面：研究大数据的表示以及大数据的计算模型及其复杂性。

(3) 在大数据应用基础理论方面：研究大数据与知识发现，大数据环境下的实验与验证方法以及大数据的安全与隐私。相比较而言，计算机科学学科是研究算法的科学，而数据科学不局限于此，其研究对象是数据，随着计算机应用从以计算为中心逐渐向以数据为中心的迁移，数据科学与工程学科的内涵和外延更加宽泛。软件工程学科中的相关技术提供了数据分析处理的工具以及具体开发时的范式。数据处理技术是数据研究领域的一种重要的研究方法，用于研究和发现数据本身的现象和规律。数据科学与工程也不同于传统的商业智能和统计学，商业智能主要从商业模式、经济管理的角度对数据应用进行研究，而统计学提供具体的数据分析处理的方法论。

1.7 大数据的挑战

大数据时代的数据存在如下几个特点：多源异构，分布广泛，动态增长，先有数据后有模式。正是这些与传统数据管理迥然不同的特点，使得大数据时代的数据管理和应用面临着新的挑战，下面将对其中的主要挑战进行详细分析。

1.7.1 大数据管理方面带来的挑战

大容量和多种类的大数据处理将带来企业信息基础设施的巨大变革，也会带来企业信息技术管理、服务、投资和信息安全治理等方面的新的挑战。如何利用公有云服务来实现企业外部数据的处理和分析？对大数据架构采用什么样的管理和投资模式？对大数据可能涉及的安全和数据隐私如何进行保护？……这些都是企业应用大数据需要面对的挑战。

1.7.2 大数据技术方面带来的挑战

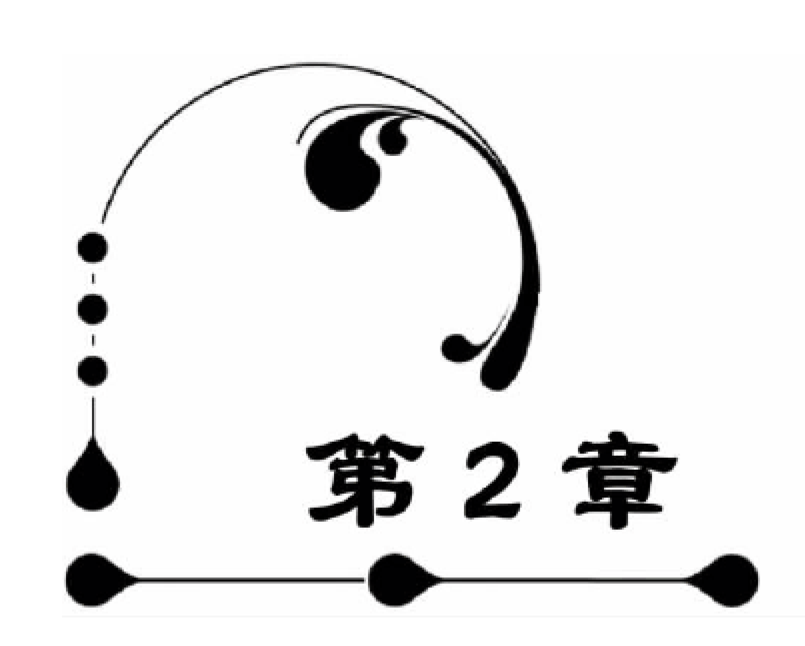
传统的关系型数据库(RDBMS)和结构化查询语言(SQL)面对大数据已经不能满足，更高性价比的数据计算与存储计算和工具不断涌现。对于已经熟练掌握和使用传统技术的企业信息技术人员来说，学习、接受和掌握它需要一个过程，从内心也会认为现在的技术和工具足够好，对新技术产生一种排斥的心理，怀疑它只是一个新的噱头。新技术本身的不成熟、复杂性和用户不友好性也会加深这种影响。但大数据时代的技术变革已经不可逆转，企业必须积极迎接这种挑战，以包容的方式迎接新技术，以集成的方式实现新老系统融合。

1.7.3 大数据工程方面带来的挑战

企业通过内部 ERP、客户关系管理(CRM)、供应链管理(SCM)、BI 等信息系统建设，建立高效的企业内部统计报表、仪表盘等决策分析工具、为企业业务敏捷决策发挥了很大作用。但是，这些数据分析只是冰山一角，这些报表和仪表盘其实是“残缺”的，更多潜在的有价值的信息被企业束之高阁。大数据时代，企业业务部门必须改变他们看数据的视角，更加重视和利用以往被放弃的交易日志、客户反馈、社交网络等数据。这种转变需要一个接受过程，但实现转变的企业则已经从中获得巨大收益。据有关统计，电子商务企业亚马逊三分之

一的收入来自大数据相似度分析的推荐系统的贡献。京东利用大数据行为分析模型缩短了电子商务时间一半以上。花旗银行新产品创新的创意很大程度来自各个渠道收集到的客户反馈数据。因此,在大数据时代,业务部门需要以新的视角来面对大数据,接受和利用好大数据,创造更大的业务价值。

大数据最根本的挑战是显而易见的:你现在的和潜在的对手总是比你更善于利用他们自己数据的潜在价值。首先他们能以更快的步伐和更小的代价重组企业,其次他们能获取实时有效的信息来制定决策,最后他们能在新产品和新市场上立于不败之地。简言之,游戏的赢家往往是那些更了解市场和消费者并根据这些信息采取行动的人。



第 2 章 大数据战略与商业模式变革

2.1 大数据战略

纵观全球大数据应用领先的国家，其产业政策大都具有以下特征：将大数据提升到国家战略层面进行布局；颁布配套产业政策扶持大数据推广；探索数据隐私法来应对数据爆炸。

信息技术与经济社会的交汇融合引发了数据迅猛增长，数据已成为国家基础性战略资源，大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。大数据是 21 世纪以来能够实现弯道超车的行业之一，各国都将其作为重要战略进行推广，主要措施有加大国家数据开放程度、扶持技术发展、推出产业扶持政策、政府立项来实现大数据的推广。随着数据的逐渐开放和数据利用程度的迅速提高，必然会面临隐私保护的难题。因此各国也在尝试制定数据隐私保护条例。如图 2-1 所示为大数据战略规划图例。

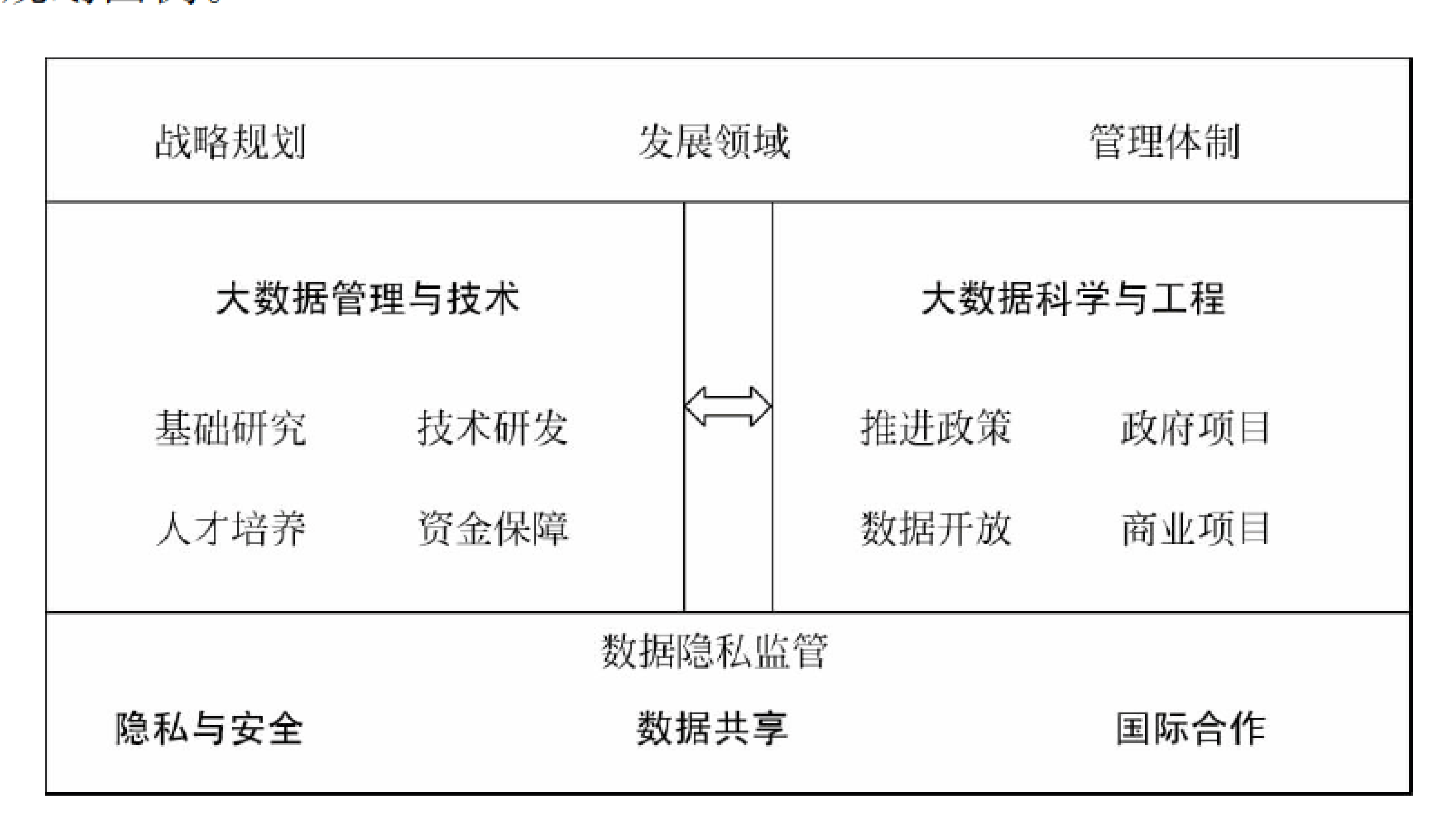


图 2-1　大数据战略规划图

“战略规划”层旨在通过分析国家级大数据战略或规划，探析西方国家发展大数据的目标定位、主要内容、重点发展的大数据应用领域，以及相应的管理体制等，总结各国大数据战略规划特色及要点。战略规划的制定为大数据技术能力储备、大数据推广应用与项目实施提供宏观指导与执行依据。

“大数据管理与技术”层探讨各国政府在大数据技术储备方面的相关政策措施，包括基

础研究部署、核心技术研发、为相关产业和研究机构提供的技术创新扶持、人才培养以及技术研发资金保障等。技术能力提升为战略规划的落地提供技术方面的支撑。

“大数据工程与科学”层从推进政策和项目实施两个角度，研究为确保大数据推广应用与项目实施而制定的各项政策，包括数据开放政策、数据共享政策、数据安全与隐私保护政策，以及政府和商业领域的试点项目规划等。数据工程为战略规划的落地提供制度支撑和实施保障。

2.1.1 国外大数据战略视角

全球已经进入大数据时代，互联网上的数据量每两年会翻一番。截至2013年，全球数量为4.3ZB，2020年有望达到40ZB。

如果将数据视为一种生产资料，大数据将是下一个创新、竞争、生产力提高的前沿，是信息时代新的财富，价值堪比石油。大数据所能带来的巨大商业价值，被认为将引领一场足以与20世纪计算机革命匹敌的巨大变革。根据IDC预测，2015年大数据市场规模将从2010年的32亿美元增长到170亿美元，复合年增长率为40%。

当前，世界各国政府和国际组织都认识到了大数据的重要作用，纷纷将开发利用大数据作为夺取新一轮竞争制高点的重要抓手，实施大数据战略。世界工业发达国家纷纷制定相关政策，积极推动大数据相关技术的研发与落实。

1. 美国大数据战略——制定计划、加强立法

2011年，总统科技顾问委员会提出建议，认为大数据具有重要战略意义，但联邦政府在大数据相关技术方面的投入不足。作为回应，美国白宫科学和技术政策办公室(OSTP)建立了大数据高级监督组以协调和扩大政府对该领域的投资，并牵头编制了《大数据研究与发展计划》(以下简称《计划》)。2012年3月29日，《计划》正式对外发布，标志着美国率先将大数据上升为国家战略。

《计划》旨在大力提升美国从海量复杂的数据集合中获取知识和洞见的能力。具体实现以下三个目标。

(1) 开发能对大量数据进行收集、存储、维护、管理、分析和共享的最先进的核心技术；

(2) 利用这些技术加快科学和工程学领域探索发现的步伐，加强国家安全，转变现有的教学方式；

(3) 扩大从事大数据技术开发和应用的人员数量。

第一波纳入《计划》的联邦政府部门主要有：国家科学基金会、国家卫生研究院、能源部、国防部、国防部高级研究计划局、地质勘探局等，投资两亿多美元，推动大数据技术研发。大数据发展不能仅靠政府，因此《计划》还鼓励产业、大学和研究机构、非盈利机构与政府一起努力，共享大数据提供的机遇。

2014年5月1日，美国白宫发布了《美国白宫：2014年全球“大数据”白皮书》，阐述了大数据带来的机遇与挑战。报告认为，大数据技术为美国经济、人民的健康和教育、国家安全、能源利用率等提供了难得的机遇，同时，报告也揭露了大数据为美国社会带来的问题，其中最重要的是个人隐私问题。

该白皮书中列举的奥巴马政府关于公开数据的举措包括政府公开数据计划、“我的大数据”计划等。政府公开数据计划为联邦数据管理工作提出了新的准则：在保护好隐私安全

性与机密性的同时，将数据公开化以及可读写化纳入政府的义务范围。“我的大数据”计划具体包括“蓝纽扣”计划、“创建副本”计划、“绿纽扣”计划、“我的学生数据”计划。“蓝纽扣”允许消费者安全地获取他们的健康信息，使得他们可以更好地管理他们的健康与经济状况，并与信息提供者交换相关信息。“创业副本”计划将纳税人的信息数据加以共享，纳税人可以通过它获得他们自己最近三年的纳税记录，这使得居民进行抵押、学生贷款、商务贷款等活动与填写纳税表更加便捷。“绿纽扣”计划为家庭与企业提供了便捷的途径来获得他们的能源使用信息以更好地管理他们的能源消耗状况来达到节约资源的目的。“我的学生数据”计划是教育部将助学金免费申请表与联邦助学情况的一些信息共享，这些信息囊括借贷、补助金、注册与超额偿付等方面的具体事项，这使得学生与资助人能够上网下载所需信息资源。在这些计划中，信息都是通过“注重使用者体验”“机器可读写”“文本信息平面化”的方式实现共享的。

此外，美国政府认为目前大数据应用中最严峻的挑战是如何保护隐私，并且正在不断修改相关法律法规以加强隐私保护，提出未来的改进重点在于：改进消费者隐私权法案、通过有关国家数据外泄的立法、保护非美籍人士隐私、规范在校学生数据采集使用、修正电子通信隐私法等。

2. 澳大利亚大数据战略

2012 年 10 月，澳大利亚政府发布《澳大利亚公共服务信息与通信技术战略 2012—2015》，强调应增强政府机构的数据分析能力从而实现更好的服务传递和更科学的决策，并将制定一份大数据战略作为战略执行计划之一。2013 年 2 月，澳大利亚政府信息管理办公室（AGIMO）成立了跨部门工作组——“大数据工作组”，启动了《公共服务大数据战略》（以下简称《战略》）制定工作，并于 2013 年 8 月正式对外发布。

《战略》以 6 条“大数据原则”为指导，旨在推动公共部门利用大数据分析进行服务改革，制定更好的公共政策，保护公民隐私。这 6 条大数据原则分别为：数据是一种国家资产，应被用于人民福祉；数据共享和大数据项目开发过程中严保用户隐私；数据完整和过程透明；政府部门间以及政府与产业间应共享技术、资源和能力；与产业和学术界广泛合作；加强政府数据开放。《战略》还决定成立数据分析卓越中心（DACOE），该中心将通过构建一个通用的能力框架帮助政府部门获得数据分析能力，并促成政府与第三方机构合作以培养分析技术专家。《战略》列举了 2014 年 7 月前需完成的 6 项大数据行动计划，分别为：制定信息资产登记簿；跟踪大数据分析的技术发展；制定大数据最佳实践指南；总结明确大数据分析面临的各种障碍；强化大数据分析的相关技术和经验；制定数据分析指南。具体工作由大数据工作组与数据分析卓越中心协作完成。

3. 英国大数据战略——做好战略布局，获取商业利益，树立政府形象

英国政府积极应对大数据时代的挑战，并且通过透明政府、智慧政府、责任政府等一系列战略布局在获取大数据带来的商业利益的同时树立开放的政府形象。

英国政府十分重视大数据的开放。早在 2012 年 12 月，英国数据战略委员会成立了世界上首个非盈利性的开放数据协会（Open Data Institute，ODI），目的就是推动开放数据的进程。ODI 是非盈利性组织，它把人们感兴趣的所有数据融会贯通在一起，每个行业的各个领域一方面产生各种数据而另一方面又可以利用这些数据。英国政府通过利用和挖掘公

开数据的商业潜力，为英国公共部门、学术机构等方面的创新发展提供“孵化环境”，同时为国家可持续发展政策提供进一步的帮助。据英国教育和科技部长戴维·威利茨介绍，ODI研究所将为那些对公众有益的商业企业活动提供数据背景支持，这将释放新的商业潜力，推动经济发展以及个人收入增长的新形式。

2013年10月31日，英国发布《把握数据带来的机遇：英国数据能力战略》。该战略由英国商业、创新与技术部牵头编制。战略旨在促进英国在数据挖掘和价值萃取中的世界领先地位，为英国公民、企业、学术机构和公共部门在信息经济条件下创造更多收益。为实现上述目标，该战略从提升数据分析技术、加强国家基础设施建设、推动研究与产业合作、确保数据被安全存取和共享等几个方面做出了部署，并做出11项行动承诺，确保战略目标得以落地。

英国政府还要求各公共部门在互联网(http://data.gov.uk)上开通开放数据的通道，向全社会开放政府管理、机构运营以及各项统计数据等相关信息。

在此基础上，英国政府发布了《开放数据白皮书》，建立了一套对公共部门开放数据程度的评价体系，对各公共部门完成开放数据任务情况进行审计，以促进英国公共服务数据的开放性。

4. 日本大数据战略——将大数据作为ICT战略重点，开发大数据应用

2012年6月，IT战略本部发布了电子政务开放数据战略草案，宣称政府将利用标准化技术生产信息确保国民方便获取数据，并保证紧急情况下以较少流量向手机用户推送信息。2012年7月，日本总务省发布《面向2020年的ICT综合战略》，重点开发大数据应用所需智能技术，创新传统IT产业，活跃“ICT的日本”。

2013年6月，安倍政府再颁新战略“创建最尖端IT国家”，阐述了2013—2020年间以开放公共数据和大数据为核心，在日本建成“世界最高水准、广泛运用信息产业技术社会”的目标。“创建最尖端IT国家”战略的要点包括：向民间开放公共数据、促进大数据的、促进个人数据的流通与运用、实现农业的知识产业化、构筑医疗信息连接网络、活用IT技术对社会基础设施进行维护管理、改革国家及地方的行政信息系统等。

日本大数据产业发展中，在个人信息保护法等法律基础设施方面落后于欧美国家，关于个人信息、保护隐私等问题，日本政府将成立研究机制针对法律措施的必要性等展开研究，修改和进一步完善个人信息保护法规也已经被提上日程。

5. 联合国——共建实验室，推动大数据解决全球问题的创新模式

2014年5月14日，联合国秘书长潘基文倡议“联合国全球脉动”发起“大数据应对气候挑战”，推动利用大数据和分析手段采取气候行动和提出创新办法，并将其中的两个获胜项目“提供森林实时信息的监控系统”和“为哥伦比亚农民推广气候智能型农业的工具”列入联合国秘书长2014年气候峰会。

2014年8月，联合国开发计划署首次携手科技企业共建大数据实验室。大数据联合实验室将利用大数据技术和联合国的全球发展经验，在环境保护、医疗与疾病预防、教育、扶贫等诸多领域进行深入的研究分析，推动大数据解决全球问题的创新模式，促进可持续发展。

发达国家大数据战略情况比较见表 2-1。

表 2-1　发达国家大数据战略比较

国家	战略规划名称	战略目标	战略内容	重点发展领域	管理体制
美国	大数据研究与发展计划	研发核心技术；推动科技进步和国家安全；培养大数据人才	牵头部门；核心项目；资金投入	科学研究；卫生；能源；国防与国家安全	白宫科学和技术政策办公室战略制定；大数据高级监督组监督执行
澳大利亚	公共服务大数据战略	推动公共部门利用大数据分析创新服务、制定最佳公共政策	未来机遇与收益；大数据应用原则；行动计划及部门分工	—	设立跨部门大数据分工组负责战略落地，成立数据分析卓越中心负责配合执行
英国	英国数据能力战略	实现英国在数据挖掘和价值萃取的世界领先地位	强化数据分析技术；加强国家基础设施建设；推动研究与产研合作；确保数据被安全存取和共享	—	英国统计局和经济社会研究委员会负责政府的数据能力提升；信息化基础设施领导理事会负责大数据基础设施建设；各行业协会负责本行业数据能力建设；信息经济委员会负责制定具体战略实施路径
法国	法国政府大数据5 项支持计划	促进本国大数据发展，推动经济社会发展	人才培养；基础设施建设；资金扶持；项目规划	人才培养；交通；医疗卫生	—

2.1.2　国内大数据战略视角

大数据产业发展是云计算技术、物联网、移动互联网迅速发展和广泛应用的结果。美国、日本、法国、韩国、澳大利亚等国家相继启动了推动大数据产业发展的政策改革，并把大数据产业发展纳入国家发展战略，通过有力的资金和政策支持加强大数据研究，优化其发展环境，抢占大数据产业发展的制高点，使其成为推动国民经济社会发展的新手段。鉴于发达国家对大数据产业的强力推动，大数据在经济、国家安全、社会、科研等方面的巨大价值和适应经济社会发展的要求，中国各级政府和社会各界也纷纷制定相关政策推动大数据产业深入发展，如表 2-2 所示。

运用大数据推动经济发展、完善社会治理、提升政府服务和监管能力正成为趋势，我国相继制定实施大数据战略性文件，大力推动大数据发展和应用。目前，我国互联网、移动互联网用户规模居全球第一，拥有丰富的数据资源和应用市场优势，大数据部分关键技术研发取得突破，涌现出一批互联网创新企业和创新应用，一些地方政府已启动大数据相关工作。坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

表 2-2 国内大数据战略进度

时间	部　　门	政策行动名称	政策行动内容
2012.7	国务院	《“十二五”国家战略性新兴产业发展规划》	明确提出支持海量数据存储、处理技术的研发和产业化
2013.1	工业和信息化部、发展改革委员会、国土资源部、国家电力监管委员会、能源局等五部委	《关于数据中心建设布局的指导意见》	建设超大型数据中心
2013.8	国务院	《关于“宽带中国”战略及实施方案的通知》	通过实施“宽带中国”战略和扩大“信息费”的方式，推动数据中心和数据中心市场发展
2013.8	国务院	《关于促进信息消费扩大内需的若干意见》	推动商业企业加快信息基础设施升级，增强信息产品供给能力，形成行业联盟，制定行业标准，构建大数据产业链，促进创新链与产业链有效嫁接
2014.8	发展改革委员会、工业和信息化部、科技部、公安部、财政部、国土资源局、住房建设部、交通运输部	《关于促进智慧城市健康发展的指导意见》	提出加强基于云计算的大数据开发与利用，在电子商务、工业设计、科学研究、交通运输等领域，创新大数据商业模式，服务城市经济社会发展
2015.3	李克强总理	2015 年中央政府工作报告	新兴产业和新兴业态是竞争高地，制定“互联网＋”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展，引导互联网企业拓展国际市场
2015.3	发展改革委员会	《创新投资管理方式建立协同监管机制的若干意见》	提出运用互联网和大数据的技术来创新监管的方式
2015.6	工业和信息化部	加快推进云计算与大数据标准体系建设	将加快云计算与物联网、移动互联网、现代制造业的融合发展与创新应用，积极培育新业态、新产业，加快推进云计算与大数据标准体系建设
2015.7	国务院	《国务院办公厅关于运用大数据加强对市场主体服务和监管的若干意见》	提高大数据运用能力，增强政府服务和监管的有效性
2015.7	工业和信息化部	将编制《大数据产业“十三五”发展规划》	大数据产业第一次明确出现在规划中还将出台促进大数据产业发展的推进计划，促进规划、标准、技术、产业、安全、应用的协同发展
2015.7	国务院	《积极推进“互联网＋”行动的指导意见》	推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展，引导互联网企业拓展国际市场

续表

时间	部　　门	政策行动名称	政策行动内容
2015.8	国务院	《促进大数据发展行动纲要》	一是打造精准治理、多方协作的社会治理新模式；二是建立运行平稳、安全高效的经济运行新机制；三是构建以人为本、惠及全民的民生服务新体系；四是开启大众创业、万众创新的创新驱动新格局；五是培育高端智能、新兴繁荣的产业发展新生态
2015.9	工业和信息化部	组织起草《大数据标准化白皮书》	制定大数据标准体系，已经开展数据质量、数据安全、数据开放共享和交易等方面的多项国家标准的立项和研制工作，同时还要积极参与 ISO/IEC、ITU 等国际标准制定工作，与国际同步发展
2015.10	十八届五中全会	《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》	实施网络强国战略，实施“互联网＋”行动计划，发展分享经济，实施国家大数据战略
2016.1	国务院	《关于组织实施促进大数据发展重大工程的通知》	提出加快落实《大数据纲要》，从破解制约大数据创新发展的突出矛盾和问题出发，重点推进数据资源开放共享，推动大数据基础设施统筹，打破数据资源壁垒，深化数据资源应用，积极培育新兴繁荣的产业发展新业态

2014 年，国务院出台的《国家新型城镇化规划（2014—2020 年）》中，强调重点扶持大数据等新一代信息技术创新应用。工业和信息化部提出了支持大数据关键技术产品的研发和产业化等具体举措。国家发展和改革委员会开展“信息化（大数据）提升政府治理能力”课题研究，并与工业和信息化部联合起草了关于促进大数据发展和应用的意见等；全国信息技术标准化技术委员会、数据中心联盟等行业机构在大数据标准和服务基础测试方面取得一定成果，如表 2-3 所示。

新一代信息技术与经济社会各领域的深度融合，引发了数据量的爆发式增长，使得数据资源成为国家重要的战略资源和核心创新要素。未来，随着我国经济发展进入新常态，大数据将在稳增长、促改革、调结构、惠民生中承担越来越重要的角色，在经济社会发展中的基础性、战略性、先导性地位也将越来越突出。同时，大数据也将重构信息技术体系和产业格局，为我国信息技术产业的发展提供巨大机遇。

欧美等国家已经出台国家战略，对其国内产业发展起到积极的推进作用。大数据发展的生态环境比较复杂，产业格局尚未形成，需要国家层面的战略规划指明大数据的发展方向、发展重点和发展路径，并处理好数据的开放、技术、应用、安全等问题。随着 2015 年 9 月《促进大数据发展行动纲要》的出台，赋予了大数据作为建设数据强国、提升政府治理能力推动经济转型升级的战略地位，国家出台各项促进大数据产业发展政策，强调产业间融合协调促进共发展，同时鼓励支持和指导地方大数据产业和应用发展，在出台产业扶持政策、开展数据共享交易、法律法规等方面成效显著。

表 2-3 各大部委对大数据的支持政策

部门	具体措施
国务院	印发《国家新型城镇化规划(2014—2020 年)》：统筹城市发展的物质资源、信息资源和智力资源利用，推动物联网、云计算、大数据等新一代信息技术创新应用
	发布《国务院关于促进云计算创新发展培育信息产业新业态的意见》：加强大数据开发与利用，实现数据资源的融合共享，推动大数据挖掘、分析、应用和服务
工业和信息化部	利用项目资金等手段进行前沿部署，支持大数据关键技术产品的研发和产业化
	推动全国信息技术标准化技术委员会开展大数据标准化的需求分析、标准体系框架研究及相关标准研制工作，并向相关国际标准化组织提交大数据研究提案
发展改革委员会	"信息化(大数据)提升政府治理能力"课题研究；大数据国家战略及发展纲要
统计局	国家统计局与浪潮、腾讯等 6 家企业合作，共同研究探讨建立大数据应用技术标准和统计标准，研究利用大数据完善补充政府统计数据，并共同开发大数据采集、处理、分析、挖掘、发布技术
全国信息技术标准化技术委员会	成立全国信息技术标准化技术委员会大数据标准工作组

未来，国家还需从法规制度入手，加强行业管理和安全保障。研究制定网络数据采集、传输、存储、使用管理的标准规范。加大对隐私信息保护、网络安全保障、跨境数据流动的管理，组织开展相关的专项检查和治理。推动和配合相关部门组织开展数据共享、开放、交易、安全等方面的立法研究工作。解决制约大数据产业发展体制机制因素和不确定性的市场因素，为产业和应用发展营造良好法规和市场环境。

2.2 大数据商业模式和商业机会

著名管理学大师彼得·德鲁克曾说过，当今企业间的竞争，不是产品的竞争，而是商业模式的竞争。Rappa(2004)认为，商业模式规定了公司在价值链中的位置，指导着公司如何赚取剩余价值；并指出商业模式明确了一个公司开展什么活动来创造价值，在价值链中如何选取上下游合作伙伴以及怎样与客户达成交易、为客户提供价值。商业模式即为企业通过产品或服务与价值链上下游主体之间建立的一种商务关系，包括公司所能为客户提供的价值、公司的内部组织结构、合作伙伴关系网络等用以实现这一价值并产生可持续盈利收入的要素。而大数据业务的商业模式就是围绕大数据资产和技术衍生出来的商业模式。大数据作为继云计算、物联网之后 IT 产业又一次颠覆性的技术变革。如何利用大数据的信息处理方式，通过收集、处理庞大而复杂的数据信息，探索并发现新的商机、对客户和市场进行新的洞察，实现业务创新和流程创新。大数据的价值必将对现代企业的管理运作理念、市场营销决策以及消费者行为模式等产生巨大影响，使得企业商务管理决策越来越依赖于数据分析而非经验甚至直觉。因而，大数据也必将对这种传统的商业模式进行近乎彻底的颠覆与模式的重构。

2.2.1 基于大数据的商业模式创新

基于"大数据"资源工具化运用的商业模式基本构成要素的创新，基本上属于熊彼特创新的范畴，它是以新资源和新技术供应为基础的产品、生产方法、市场及行业的转变；这种

创新是建立在新的数据资源观基础之上的，它包括对“大数据”资源本身价值、利用方式、获得方式的再思考，也包括对受“大数据”影响的企业其他资源、能力延伸和利用方式的再思考。基于“大数据”的企业特征层面的商业模式创新主要表现为：价值主张创新、价值创造和传递模式创新（关键业务和流程创新）、收益模式创新，以及外部关系网络和价值网络重构。

1. 基于“大数据”的价值主张创新

“大数据”由于具有无限接近消费者的潜能可以为企业提供精准的价值主张。

（1）洞悉消费者的真实需求。面向顾客的公司很长时间以来都在利用数据细分和定位它们的顾客，然而消费者的真实需求具有隐蔽性、复杂性、易变性和情景依赖性，利用历史的、静态的、结构化的数据，企业很难获得用户的真实需求。而“大数据”使企业获得消费者的真实需求成为可能：人类的细微行为，会直接暴露内心的真实想法，例如网友在网络中的足迹、点击、浏览、留言等能直接反映他的性格、偏好、意愿；在物联网世界，企业可以运用来自内置于产品中的传感器数据，了解商品在真实世界里的真实使用情况。

（2）对消费者进行准确细分。传统的、企业可操作的消费者细分一般以地理位置、人口统计特征为依据，而“大数据”可以实现越来越接近消费者真实需求的细分方式：一是细分标准抽象化，当人们的兴趣、爱好、价值观、生活方式、沟通方式等都可以数据化以后，以这些特征细分消费者就具有了现实可行性；二是细分市场微小化，从本质上讲，世界上有多少人就有多少种兴趣、偏好和需求，每个人都是一个细分市场，“大数据”正在使企业向“微市场”（Micro-Segments）（Goyaletal，2012）化迈进。例如在医疗行业，基于包括个人遗传基因及分子组成的大数据的个性化医疗已经成为这一行业商业模式变革的大趋势。

（3）产品的即时、精准、动态定位。

大数据的实时个性化（Real-time Personalization）以及多来源、多格式数据的快速综合对比分析能力使数据的收集、整理、分析、反馈、响应可以在瞬间完成，使企业随时随地精准圈定用户群并满足他们的真实需求和潜在需求成为可能。零售业就是一个典型的数据驱动定制化的行业，目前在线零售商利用实时数据提供精准的商品推介已经十分普遍；新一代的零售商已经可以通过互联网点击流跟踪消费者的个人行为，更新他们的偏好、实时模型化他们的行为模式，快速识别出消费者在什么时候接近购买决策，然后打包首选商品促进交易的完成。以 Sears Holdings（希尔斯控股）为例，几年前这一公司就决定利用它的三大品牌收集关于顾客、产品、促销的巨大数据创造价值——用以量身定做针对顾客的个性化促销手段和产品。但是，这一大规模分析所需要的数据是海量的而且是碎片化的，存储在不同品牌所持有的多个数据库和数据仓库中，运用企业原有的 IT 架构完成一轮分析需要 8 个星期的时间，这使其没有商业价值。后来，公司转向了大数据技术和实践，与 Cloudera 公司合作搭建了 Hadoop Cluster（分布式计算集群），运用集群可以直接进行数据分析，避免了耗费大量时间从不同来源抽取数据加以合并才能用于分析的复杂过程，产生一套有效的促销设计的时间从 8 个星期缩减为一个星期，而且 Hadoop Cluster 的存储和运行成本仅仅是传统标准数据库成本的一个零头（McAfee，Brynjolfsson，2012），大数据技术的运用使其“量身定做”的价值主张得以实现。

2. 基于“大数据”的关键业务和流程创新

作为基础技术条件和工具，“大数据”资源具有释放和放大其他资源价值的能量。基于

"大数据"的关键业务和关键流程创新就是企业业务活动的"大数据"化，依据其改造和影响的范围可以分成以下几种情况。

（1）以"大数据"设施和技术作为基础，以数据信息流为线索对整个业务流程进行再造，例如，"大规模定制"生产方式的实现就是基于强大的 IT 基础设施对企业进行流程再造的结果。

（2）以"大数据"活动取代传统的业务流程，使企业的业务经营模式发生变化。例如，电子商务的发展就是传统商业流通主要交易流程被数据交换取代的结果。

（3）把"大数据"活动纳入价值创造流程，寻找新的价值创造方向和路径。例如在汽车行业，利用大数据分析，充分挖掘数据信息背后所隐含的行业技术关联，寻找有效途径延长燃气涡轮、喷气式发动机和其他重型设备的运行时间，这为传统制造业寻找新的价值增长点提供了思路。

（4）基于"大数据"的流程再设计，以"大数据"作为解决问题的新方法，提高某一业务流程的效率或效果。以机场为例，预计航班到达时间是机场的一个重要流程，以往这一估计由飞行员到达最后一个导航点至机场期间提供，高误差率和大的误差范围引发了相当可观的成本，利用 PASSUR Aero space 公司提供的名为 Right ETA 的航班到达时间估计服务彻底改变了这种状况：Right ETA 服务是基于天气、航行时间表等公共数据以及 PASSUR 收集的多维历史数据进行的精细分析和模式匹配分析，转向使用 Right ETA 服务以后，机场从根本上消除了预测误差，每年可为机场创造几百万美元的价值。

3. 基于"大数据"的盈利模式创新

许多商业模式创新都是建立在这样一种认知基础之上的：消费者对商品需求的本质是使用商品而非拥有商品本身。例如，出售模式改为出租模式，与此相对应的收益模式从一次性支付向"微支付"转变：著名的建筑设备制造销售商喜利得(Hilti)变身成为设备出租服务合同管理商(Johnsonetal，2008)，国内的"北森测评"公司通过在线销售创新软件收费方案——由原来以企业为单位的固定收费转变为按照使用次数收费等，这些创新都取得了巨大的成功。但是，使用这一收费模式的前提是使用过程可被记录和量化，而"大数据"可以实现使用过程、频率、强度的实时监控和记录。这一收益模式变革在软件行业和媒体广告行业最为典型。在软件行业，应用软件泛互联网化改变了消费者获得和使用软件的方式，软件价值的载体虚拟化，使软件的价值传递方式和收益模式必须发生改变。例如开源软件模式、AppStore 模式等，企业利用"门户化"建立排他性，提高客户黏性；利用"碎片化"，把原来大型臃肿的软件，拆分成多个独立的功能组件，用户可以按需下载，从而降低了客户的总体拥有成本，企业的关键流程也由开发、复制、销售软件向开发、服务、提供问题解决方案转变。在媒体广告行业，传统的以呈现时间或者频次为计费标准的收费模式很难在广告费用和广告效果之间建立起直接的联系，对于广告主来说，如何确定广告的有效性是最大的困扰，正如百货行业巨子约翰·沃纳梅克(John Wanamaker)所说：他花在广告上的钱有一半是浪费的，但却不知道是哪一半。利用"大数据"，互联网广告正在逐步实现广告成本与广告价值的对等。例如，CPC(Cost per Click)模式，即广告主为每次点击付费；CPM(Cost per Thousand Impressions)模式，即广告主以广告显示每 1000 次为单位付费；CPA(Cost per Action)模式，即广告主为广告所带来的用户的每次特定行为付费，包括形成一次交易、获得一个注册用户、产生一次下载行为等；CPS(Cost per Sale)模式，即基于广告引入用户所产

生的成功销售而收取一定比例的佣金，典型的如 Google 地图的"点击呼叫"(Click-to-Call)功能，以及 Facebook 刚刚宣布推出的"转化追踪"服务，这些创新与应用正在引发广告媒体行业收益模式的大变革。

4. 基于"大数据"的关系网络和价值网络重构

从 RBV 资源分析视角看，数据资源虽然具有很高的价值，但是其流动性强、可获得性强、价值流逝速度快而且对它的利用方式也易于模仿，而且它还具有无形性、知识性特征。大数据技术具有高度专业性和复杂性，大数据基础设施的运行具有高固定成本、低边际成本的特征，而且对其访问(利用)呈现高度并发性和波动性，企业以传统方式获取和控制大数据资源和技术成本高昂，而且风险很大；而另一方面大数据技术却使外部资源利用的交易成本和风险大大降低，这就使得企业在"大数据"资源获得和利用方面倾向于选择介于市场交易与内部生产之间的方式，分享与合作成为企业构建外部关系网络和价值网络的主题；例如数据共享、IT 外包等，IT 外包是目前一般企业解决"大数据"问题的基本思路，也是"大数据"产业链形成的根本推动力，这一方式可以实现"大数据"资源的柔性配置和规模效率。

除了获取大数据资源和技术本身为目的的外部合作以外，大数据技术使企业获取和利用其他外部资源的成本和风险也大大降低，为新的价值创造模式和价值传递模式提供了技术路径。

(1) 众包(Crowdsourcing)。众包是指把传统上由指定代理人(通常是雇员)完成的任务以公开选拔的形式外包给大量不特定的个人去做的行为(Howe，2006)。众包模式的实质是对离散社会资源的有效利用。在 IT 业，开源社区(Open Source Community)就是众包的典型模式，目前各大 IT 巨头都争相采取这种模式构筑自己的创新"生态圈"，其他行业的许多世界性大公司也都建立了自己的网络平台或者借助众包中介(Crowd sourcing Intermediates)以众包方式解决技术、创意、设计等原来完全由内部流程和资源完成的活动，如宝洁、杜邦、波音等。

(2) 用户自生成内容(User-generated Content)。用户自生成内容是在"去中心化"、用户参与、用户体验、协同创作等互联网文化推动下产生的一种新兴的网络信息资源创作与组织模式(赵宇翔等，2011)，消费者以上传文字、图片、音频、视频或者共享文件等形式参与内容和价值创造，这一模式的典型代表如维基百科、Google、Facebook 等。

(3) 共同创造(Co-creation)。从比较深层的意义上看，共同创造是把消费者、供应链成员乃至其他相关产品提供者纳入产品价值网络的思维方式。从简单意义上看，是指企业整合来自于多元系统的数据、邀请跨职能部门的合作甚至从外部供应商和消费者那里获取信息以共创产品(Leeetal，2012)。例如，汽车行业基于集成化数据平台的全供应链设计合作，玩具行业巨头乐高基于在线订购的允许客户组装他们自己乐高套件的乐高工厂等。这些新模式所依赖的核心工具都是基于 Web 3.0 技术的网络平台。

这些创新改变了企业对外部资源需求的内容及方式，改变了企业创造价值、传递价值的方式及路径，改变了企业的商业生态，使企业的资源边界、市场边界和契约边界都呈现模糊化趋势。可见，企业对"大数据"资源的获得和利用过程也是企业重构外部关系网络和价值网络的过程，价值网络重构已经成为企业商业模式创新的重要方式之一(王琴，2011)。

信息资源产品化的基本前提是信息的可分离性(Information Separability)，即各种无形的信息能在多大程度上以数字的形式被捕捉从而与产生它的活动相分离，使其可以用来指

导下一次活动(Sampler,1998)。“大数据”的发展为信息的分离提供了载体和工具:用户在各类信息平台上留下了海量数据,在大数据处理技术之下可以对其进行分类整理和重新聚合,这些聚合性的数据信息包含着极高的商业价值,并且具备了销售的可能,至此数据信息得以向数据产品过渡(黄升民,刘珊,2012),以“大数据”产品为核心的产业链正在形成。“大数据”产业链可以从两个方向进行描述:以大数据产品价值链为线索沿横向从数据采集、整理、分析到决策逐级递进,以大数据技术为中心沿纵向从底层的基础设施供应、大数据技术提供到完整 IT 解决方案服务。从产业价值链的层面看,不同的商业模式主要是指企业在产业链上的不同角色和地位,商业模式创新则来自于企业在价值链上的重新定位、价值链的延展、分拆、创新与混合(高闯,关鑫,2006),这一层面商业模式创新的基本趋势是以产品为中心的价值链定位与选择正在向满足客户完整解决方案需求的业务活动选择的转变,从而使“大数据”产业呈现与其他产业交叉重叠的趋势。

(1) 企业价值链水平延伸商业模式。

在“大数据”行业,按照加工深度的不同,数据产品基本上可以分为数据(原始数据)、信息和知识。数据(Data)是载荷或记录信息的按一定规则排列组合的物理符号,可以是数字、文字、图像,也可以是计算机代码。拥有数据是获取信息的第一步,信息的获取还需要对数据背景进行解读,即当接收者了解了物理符号序列的规律,并知道每个符号和符号组合的指向性目标或含义时,才可以获得一组数据所载荷的信息(可以用公式“数据+背景=信息”表示),也可以说,信息是指把数据放置在一定的背景下,对数字进行解释并赋予意义。在此基础上,使用者通过对这些数据的转换、整合、计算、分析来解释各种现象背后的原因,预测事物的发展趋势,并应用于具体的专业实践活动,数据就成了“知识”(黄升民,刘珊,2012)。大数据产品的价值取决于数据资源的专有性(Data Specificity)程度,即数据资源的使用或获得在多大程度上限定于特定的个人或者特定的时间期限,其中,个人专有性也称为知识专有性,是指只有拥有特定知识的人才能获得或使用,也就是其获得和利用对某种特定知识的要求;时间专有性是指数据资源必须在其产生后的很短时间立即被捕捉,必须在其产生后的特定时间段内被使用。数据、信息、知识的获得时间专有性和获得知识专有性程度不同,也就决定了其价值创造所依赖的关键资源不同,从而也就决定了拥有不同核心资源和能力的企业在价值链上的不同定位。

基于此,以数据产品为基本提供物的数据公司,按照其在大数据产品价值链上的不同定位,可以分为三种基本商业模式:

① 数据租售模式。这一模式的价值主张是向客户提供原始数据的租售,其关键流程是数据的采集、传输和整理。原始数据的获得时间专有性很强,也就是必须要有实时接触和采集数据的条件,但其获得知识专有性相对较弱,所以,这一商业模式所依赖的核心资源是有利的采集数据的技术基础和条件。这一商业模式处于价值链第一阶段。例如,2010 年在深圳中小板上市的四维图新公司,其价值主张是以覆盖全国的高质量导航电子地图数据库及其更新体系满足汽车工业、消费类电子行业、互联网和移动位置服务等各行所需。它处于产业链最上游,精准的导航数据是公司的核心产品,也是地理信息数据及应用产业最稀缺的资源,这家公司因此成为国内第一家上市的导航电子地图生产企业。

② 信息租售模式。这一模式的价值主张是向客户提供代表某种主题的相关数据集,诸如数据包租售等,其关键流程是把原始数据与其背景意义相结合,整合、提炼、萃取,使数据

形成价值密度更高的信息。信息的获得时间专有性相对不强，但其获得知识专有性较强（主要是数据处理领域的知识），所以，这种商业模式所依赖的核心资源是数据处理技术及能力，这种商业模式处于价值链的中间阶段。例如彭博（Bloomberg）公司，其价值主张是为专业人士提供及时、准确、丰富的金融交易信息和财经资讯，公司的核心竞争力在于积累了丰富、大量的金融行业数据和交易数据，拥有强大的专家和咨询网络，构建了整合专业服务与媒体服务的全球性服务平台，彭博也因此成为全球商业、金融信息和财经资讯的领先提供商。

③ 知识租售模式。这一模式的价值主张是为客户提供一体化的业务问题解决方案，其关键流程是将“大数据”与行业知识利用相结合，通过行业专家，深度介入客户的业务流程，提供业务问题解决方案。相对而言，知识的获得时间专有性较弱，但其获得知识专有性很强（包括数据处理知识和特定行业知识），所以，这一商业模式所依赖的核心资源是拥有大数据挖掘技术的行业专家，这种模式实际上已经超越了数据公司的范畴。例如 Opera 公司，它致力于提供大数据的挖掘，在高度专业化的领域提供高端的服务，其业务诸如：为银行信用卡部门设计新的产品和营销方案，帮助保险部门确定寿险、车险等的赔率，帮助投行确定应该对哪些用户推出新的产品，等等。可以看出，这种模式已经具有了跨行业的特征。

(2) 大数据为中心垂直衍生商业模式。

广义的大数据技术可分成 4 个层面：平台层（并行构架和资源平台，即硬件层面）、系统层面（大数据存储管理和并行编程模型与计算框架）、算法层（基础算法和应用算法）和应用层（应用开发和行业应用）（黄宜华，2012）。狭义的大数据技术则仅包括后三个层面（即软件层面）。在“大数据”行业，以大数据技术为基本提供物的大数据技术公司，它们为其他行业企业以及数据公司提供 IT 基础及服务，按照其在大数据技术纵向架构中的不同定位，可以分为三种基本商业模式，即硬件租售模式、软件租售模式和服务模式，服务模式已经成为这一领域商业模式创新的大趋势。

① 硬件租售模式。采用这一模式的企业主要包括大数据存储设施、计算设施、网络设施的销售商，也包括新兴的提供云存储、云计算业务的服务提供商（相当于硬件设施的出租）等，Dropbox、国内的微盘、华为、联想都是此类公司的代表。例如，Drop Box 就是 Dropbox 公司运行的在线存储服务，通过云计算实现因特网上的文件同步，用户可以存储并共享文件和文件夹，采取免费＋收费的商业模式，它为初始用户提供 2GB 的免费文件空间，用户可以通过邀请其他人参与、使用以及付费等方式获得更多文件空间。

② 软件租售模式。采用这一模式的企业主要是指大数据技术（狭义）与服务提供商，这些提供商围绕 Hadoop 架构开展一系列研发，提供大数据存储、检索、数据挖掘等技术和服务，它们提供专为解决数据挑战而创建的优化型技术，用以捕获、处理、分析和显示非结构化和结构化数据，并将其转换为有意义的洞察性信息。例如在算法层面，目前国内提供非结构化数据处理技术的代表性公司有：语音数据处理领域的科大讯飞，视频数据处理领域的捷成股份，语义识别领域的拓尔思，图像数据处理领域的超图软件，大数据存储领域的同有科技公司，等等。在应用层面，例如全球商业智能和分析软件与服务领袖——SAS 公司，它一直致力于数据统计软件的开发和销售，SAS 在综合的企业智能平台上提供一流的数据整合、存储、分析和商业智能应用，帮助企业更快、更准确地进行业务决策。

③ 服务模式。这一模式建立在“大数据”行业垂直整合的基础上，需要企业与客户进行深度合作，其价值主张是为客户提供一体化的 IT 问题解决方案。“大数据”时代开源软件

的兴起和繁荣使传统的操作系统、中间件、数据库等平台级软件的同质化趋势渐趋明显，使最终用户关注的焦点转变为如何解决企业的业务问题，而不是购买谁的设备、使用谁的数据库或者操作系统，深度定制化成为需求的基本特征。在这一背景下，各大IT巨头开始通过收购、合作、创新、调整来布局自己的“大数据”业务，逐步由硬件供应、软件供应向服务模式转型，其典型代表如IBM、EMC、Oracle、SAP等。IBM在1992年开始由硬件供应商向服务提供商转变的商业模式创新，提出为用户提供完整解决方案的价值主张。面对“大数据”的到来，应对“感知化、互联化、智能化”的科技大势，又提出“智慧地球”的愿景，部署自己的“大数据”战略（Weed，2012），通过收购Cognos、ILOG、SPSS、Netezza、Coremetrics等使公司的业务涵盖企业的文化战略咨询、组织流程梳理、IT治理、系统建设、基本应用软件、中间件、数据库、操作系统、主机等，实现了向服务模式的转型。EMC通过系统、软件和服务的组合，自上而下设计、构建总成解决方案，帮助IT部门以更敏捷、更可信、成本更低、效率更高的方式存储、管理、保护、分析他们最重要的资产——信息；通过并购VMware、RSA、DataDomain、Greenplum、Isilon等多家在“云和大数据”方面具有高度战略价值的公司使公司的业务涵盖：云基础架构转型服务，关键应用程序转型服务，利用云计算实现业务转型服务等。Oracle公司在数据库产品取得行业领袖位置以后，首先向产业链下游扩张，加强对终端客户的掌控；然后向产业链上游扩张，涉足中间件供应和服务器制造，从而实现了产业链上下游的全覆盖：打包主机、操作系统、数据库、中间件、应用软件，形成战略性的新产品ExaData（“新一代海量关系数据管理平台”）（Billings，2012）。SAP在2012 SAP全球技术研发者大会上正式宣布推出基于HANA（高性能分析应用软件）平台的Business One解决方案，至此，通过与芯片、系统厂商的深度定制与紧密捆绑实现了SAP的软硬一体化战略。

这些创新源于不同的起点、沿用了不同的路径、依托不同的资源和优势，但是，它们创新的逻辑起点却是相同的：提供最佳客户体验，并在这一思想指导下实现了突破产品边界、业务边界甚至产业边界的创新。

2.2.2 大数据对企业管理决策的影响

管理的重要职能就是决策，通过决策实现管理目标，优化资源配置。管理是由一系列的决策组成的，决策需要依靠准确、完整、及时的信息和数据，在此基础上寻找优化方案，或者满意的解决方案（参考时间、机会成本等）。大数据从支持决策，再进一步到在某些领域产生决策；因为大数据的5V特征，使得数据来源维度拓展，包括领域决策的模型和数据，也包括决策主体的行为模型和数据，以及决策环境的模型和数据。大数据使得决策过程得以在更多维度空间下优化，不仅包括领域的、微观和技术的过程，也包括主体的、行为的和偏好的特征，还考虑了环境的、宏观的和系统的因素。大数据相关技术对于数据维度的扩张，数据处理能力的提升，数据层次的提升的同时，也扩大了数据的收益。

企业绩效分析和预测是企业大数据的重要应用之一。企业数据从内部ERP系统、业务系统、办公自动化系统、客户服务系统的客户反馈、员工日志、生产制造系统中获取。财务价值数据主要来自ERP、财务系统、预算系统等，也能从上市公司年报、政府统计数据、行业年鉴等中获取有用信息。客户数据主要来自内部CRM系统、呼叫中心、门户网站、社交媒体等。通过这些数据分析能够分析和预测企业业务和管理绩效，为企业运行提供全面的洞察力。按照企业平衡记分卡的模型，企业绩效的数据主要包括4个方面，分别是企业业务运营

数据、财务价值数据、客户数据和面向企业未来发展的数据。

大数据也可以帮助企业提升资产管理和优化业务流程，提升企业管理绩效。企业利用实时数据能够实现预测性的维护并减少故障，推动产品和服务开发。大数据的精髓就是它力图追求全样本，大数据的使用也存在网络效用，用户越多增值越快，通过基于全样本的检索与推断，能够完成精准的个体推荐，让企业整个生产效率极大提高。大数据系统与人的配合使得可以更好地提供答案，大数据支持模型化，支持决策的优化，也在支持问题、方法和答案匹配的优化。

UPS 快递高效地利用了地理定位数据。为了使企业总部能在车辆出现晚点的时候跟踪到车辆的位置和预防引擎故障，它的货车上装有传感器、无线适配器和 GPS。同时，这些设备也方便了公司监督、管理员工并优化行车线路。UPS 为货车定制的最佳行车路径是根据过去的行车经验总结而来的。2011 年，UPS 的驾驶员少跑了近 4828 万千米的路程。

DHL 是全球知名的邮递和物流公司。它是一家传统行业的企业，然而在移动互联网和大数据浪潮中并不落后，在瑞典推出了众包模式送货的移动应用 MyWays，人们可以通过移动应用报名投递自己行动路线附近的包裹，并获取报酬。此外，DHL 还把大数据应用于管理物流风险，从而为客户提供更好的服务。

面向企业未来发展的数据来自企业社区、知识管理、人力资源管理、企业即时通信、企业微博等系统，也有来自社会公益组织、政府的数据。通过企业内外部数据的采集和分析，能够实时反映企业战略目标的执行情况、差距，并对未来战略目标的实现进行提前的预测和分析，如图 2-2 所示。

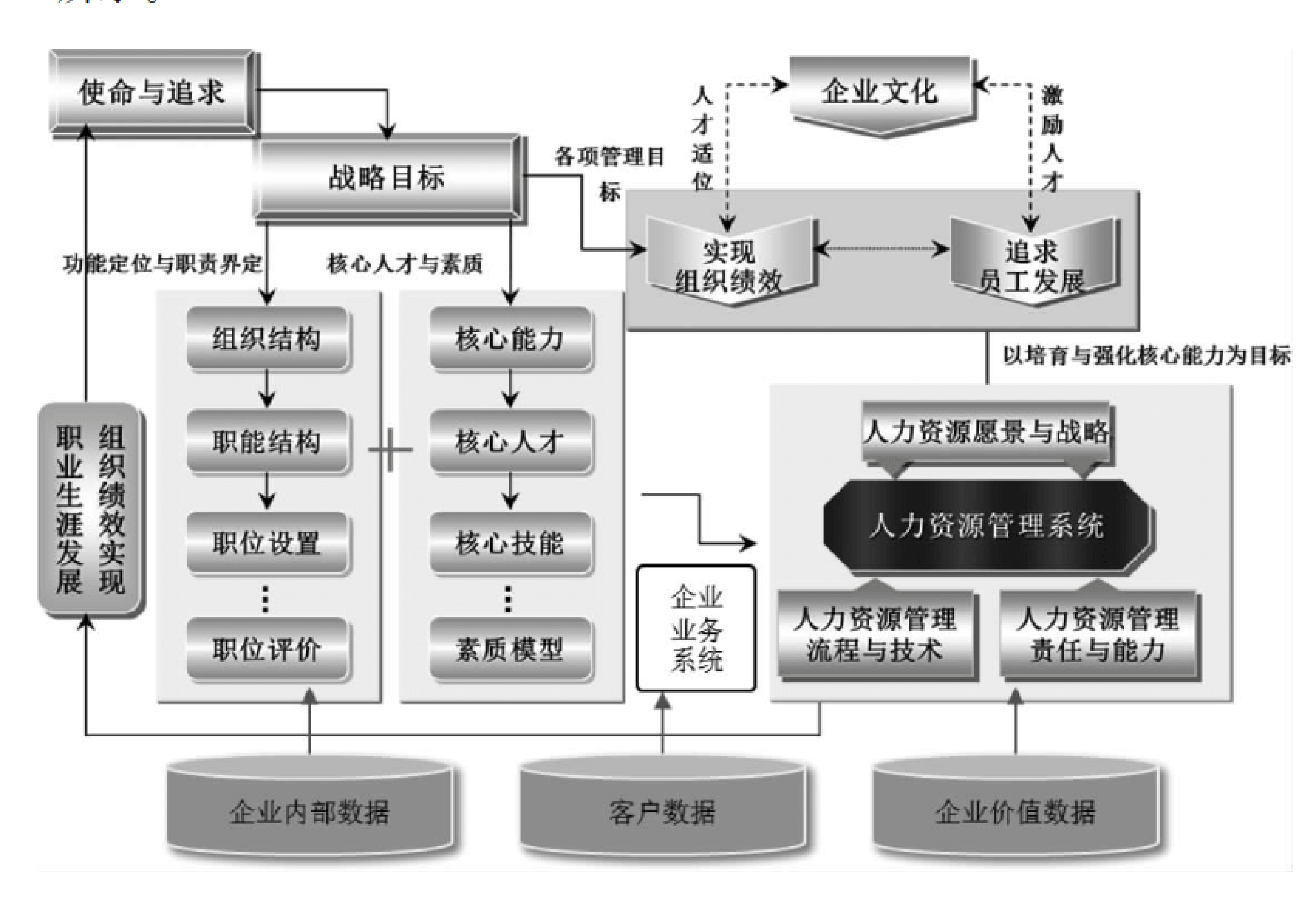


图 2-2　数据驱动的企业绩效管理

2.2.3　基于大数据驱动的商业机会

大数据驱动的商业机会不胜枚举，最经典的案例应该是美国沃尔玛公司（WalMart）将尿不湿和啤酒摆放在一起的销售策略。沃尔玛对顾客的购物习惯进行关联规则分析，从中判断顾客会经常一起购买哪些商品。沃尔玛利用数据挖掘工具对其保存在数据仓库里面的所有

门店的交易数据进行分析，得出了和尿不湿一起购买最多的商品是啤酒的结论。沃尔玛在所有的门店里将尿不湿与啤酒并排摆放在一起，结果是尿不湿与啤酒的销售量双双增长。

另外一个比较著名的例子就是 Target 怀孕预测的案例。他们对商品数据库里的数万类商品和女性顾客的商品购买记录进行分析，挖掘出与怀孕高度相关的 25 项商品，制作“怀孕预测”指数，可以精确地预测到客户在什么时候想要小孩，推算出孕妇的预产期等，从而抢先一步给女性推荐相关的产品。

通常，利用大数据进行商业机会分析，要遵循以下规则。

1. 以客户为中心，挖掘客户需求，进行销售预测

大数据在用户行为分析和预测方面的应用比较典型。通过对用户社交网站的行为数据、浏览器的日志信息、传感器的数据等进行收集和分析，就可以得到用户的行为习惯，通过建立数据模型，可以对用户的下一步行为进行预测。

例如，美国统计学家内特·西尔弗建立统计模型，成功预测了 2012 年美国大选的结果。通过他的预测，看到奥巴马有 431 种胜利途径，对比罗姆尼仅有 76 种，奥巴马总统连任的机会是 86.3%。在其他行业，电信可以通过大数据预测用户的流失，从而可以提前采取相应的手段留住客户；汽车保险行业可以了解客户的驾驶水平和需求，来为顾客推荐合适的保险等。大数据对于当代企业能够更好地运营所体现出的价值已经不言而喻。

(1) 从全局角度出发挖掘客户需求。众所周知，数据蕴含着巨大的价值。但是什么样的数据最有价值？是客户实际购买的信息？是他们心中想要的东西？他们在寻找什么？他们社交网站的活动记录？他们网上浏览时留下的记录？品牌商观察顾客的数据是否比顾客自己提供的数据更可靠？市场分析人员采用的算法和分析结果的作用是什么？当然有一种数据来源是最可靠的，“银色的数据弹”对吗?

决策者为了了解事实真相，必须从多个角度考虑问题，从不同消费者需求的角度出发，对事物要有个全局的认识。因为没有单独一种数据能够描述和预测消费者的所有行为，正如盲人摸象一样。我们分析一下原因，过去的消费记录是很重要，但即便最忠诚的顾客也会在其他品牌上花费时间。另外，实际上一个品牌最合适的消费者可能是另一个竞争品牌更棒的消费者。知道顾客在没有选择你的品牌时还有哪些行为，这会给你带来许多线索，你就可以提高顾客的份额比重并帮助创造新的产品和服务。

通常情况下，我们会询问顾客想买什么，但这并不全面。有时候消费者并不总是知道他们实际需要什么，就像是这些消费者在看到 iPod 之前有几个真的想过要买一台？另外，消费者们总是对世界抱有一些不切实际的想法，一个从数学角度来看不可能的现象是 63%的美国人认为他们拥有超过平均水平的智商。由此将消费者的想法和实际行为联系起来至关重要。

对于产品搜索是不是最好的指标？从统计学的角度而言，这个结果是难以一概而论的。搜索是非常有效的手段，但是并不能制造需求。消费者的网络活动记录是否可靠？虽然他们在完成诸如消费意愿市场调查时拥有巨大的潜力，但是和其他新兴技术一样，市场分析人员仍然在摸索如何有效地使用这些工具。为了充分挖掘市场蕴含的潜力，当今企业必须培养和管理针对消费者的多元数据分析能力。

这项战略需要综合统筹线上、线下和反映消费意愿的数据，以及通过观察、推导、自愿收集和预测等方法获取的数据，同时最大限度地激活、评估和利用这些数据。这意味着需要将

多元数据看作是企业的一项宝贵资产，将它从传统消费者分析的桎梏中解放出来。

实行多元分析就意味着需要在企业的管理层设立战略委员会来判断市场分析的成功与否。成功的策划活动或是完善的销售渠道固然不错，但是获得成功的关键在于拓展客户所带来的价值。为了实现这个目标需要研发新的科学技术来最大规模地激活和评估企业数据，我们称之为企业数据管理系统。该系统的独特之处在于可以持续执行数据分析从而和目标客户群建立高效的沟通机制。

(2) 客户特征分析。在各个行业中，大数据业务应用需求集中于满足以客户为中心的目标实现，客户分析是大数据应用的重要领域。企业希望大数据技术有能力更好地了解和预测客户行为，并能够改善客户体验。客户分析的重点是收集和分析交易数据、多渠道交互数据、社交媒体数据、会员卡服务数据及其他与客户相关的数据，以全面提高企业了解客户偏好和需求的能力，真正帮助营销、销售和客户服务部门实现客户关怀的目标。

客户分析的主要维度：一是全面的客户数据分析；二是全生命周期的客户行为数据分析；三是能为客户提供的服务价值分析。

① 全面的客户数据——客户是谁？

建立全面统一的客户信息资料，通过客户唯一的身份标识号，可以获取客户各种相关数据，包括相关业务交易和服务数据。

② 全面生命周期的客户行为信息——客户的真实需求是什么？

对于客户的历史交易、相关信息进行跟踪分析，分析客户行为特点、需要偏好，建立客户模型（比如阿里巴巴的量子恒道、数据魔方和生意参数），挖掘客户的真实需求和潜在需求。

③ 能为客户提供的产品和服务——服务价值分析？

通过分析客户的真实需求和潜在需求，让客户参与产品和服务创新，促进企业服务的改进和创新。

2. 建立全方位客户数据分析模型

(1) 客户全面基本信息模型。客户按照类型可以分为个人客户和企业客户，对应客户的基本信息不同，如个人客户记录姓名、年龄、家庭地址等数据，企业客户记录企业名称、企业注册地、企业法人等数据。从共同的属性来看，有客户基本属性和派生属性，基本属性有客户号、客户类型、客户信用度等，派生属性是由基本属性衍生分析出来的数据，如客户满意度、贡献度、风险度等。客户数据和客户交易数据、客户行为数据、客户需求数据相关联，这种关联关系是通过客户服务的交易、购买的产品、产品厂商、账户等数据来建立的，如图 2-3 所示。

(2) 客户价值需求模型。菲利普·科特勒的客户让渡价值理论。按照菲利普·科特勒的观点，顾客让渡价值（Customer Delivered Value，CDV）是指总顾客价值（Total Customer Value，TCV）与总顾客成本（Total Customer Cost，TCC）之差。总顾客价值是指顾客期望从某一特定产品或服务中获得的利益的总和，包括产品价值、服务价值、人员价值和形象价值。总顾客成本是指顾客为购买和使用某一特定产品或服务而付出的代价的总和，包括货币成本、时间成本、精力成本和体力成本（菲利普·科特勒著，梅汝和等译，2001）。

格隆罗斯的客户价值过程理论。格隆罗斯是从关系营销的角度阐述客户价值的，他认为，价值过程是关系营销的起点和终点，关系营销应该为客户和其他各方创造出比单纯交易营销更大的价值，并且必须让客户感知到持续关系中所创造的价值。

价值取向（Value Orientation）是价值哲学的重要范畴，它指的是一定主体基于自己的

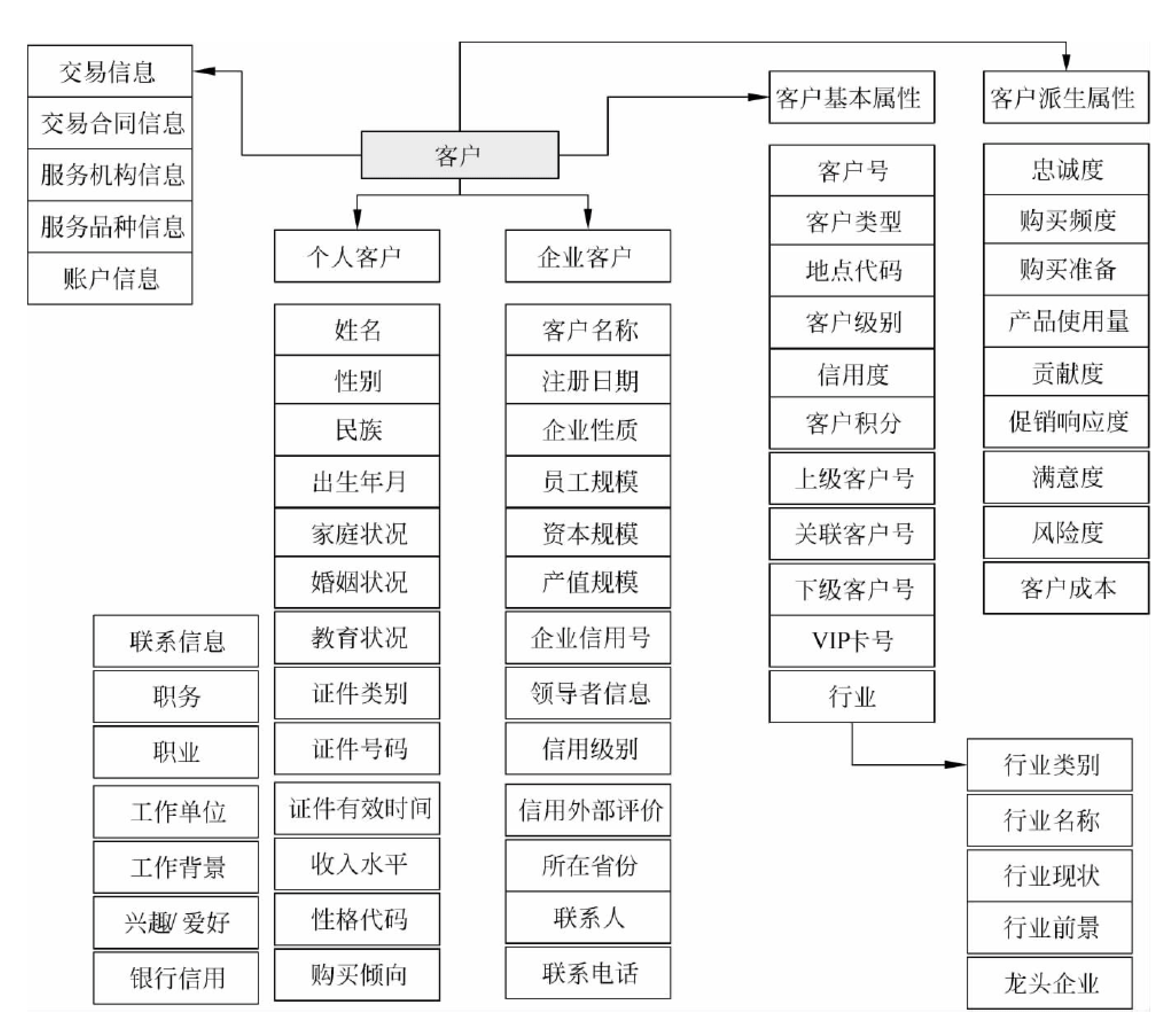


图 2-3　客户数据模型

价值观在面对或处理各种矛盾、冲突、关系时所持的基本价值立场、价值态度以及所表现出来的基本价值倾向。价值取向具有实践品格，它的突出作用是决定、支配主体的价值选择，因而对主体自身、主体间关系、其他主体均有重大的影响。人们在工作中的各种决策判断和行为都有一定的指导思想和价值前提。管理心理学把价值取向定义为“在多种工作情景中指导人们行动和决策判断的总体信念”。

人的价值取向直接影响着工作态度和行为，如图 2-4 所示为客户价值取向的决定因素模型。诺贝尔经济学奖获得者、著名心理学家西蒙认为，决策判断有两种前提：价值前提和事实前提。说明价值取向的重要性。客户价值取向是客户基于自身的价值观、需求、偏好和财务资源，在面对或处理与供应商各种矛盾、冲突和关系时所持的基本价值立场、价值态度以及所表现出来的基本价值倾向。

3. 通过客户管理策略进行数据分析预测和精准营销

(1) 遵循顾客至上销售策略。经典精英理论创始人维弗雷多．帕雷托(Vilfredo Pareto)名言指出：顶层 20%的顾客创造了约 80%的总利润。经验也告诉我们，高端客户可以比普通客户创造 5 倍甚至是 10 倍的价值是司空见惯的事。而许多企业依然没有对顾客利润测算在市场分析中的作用给予充分的重视。但现实情况是市场分析人员往往没有在拓展客户价值方面下足功夫。事实上，高达 60%的企业投入了仅仅 20%甚至更少的市场运作

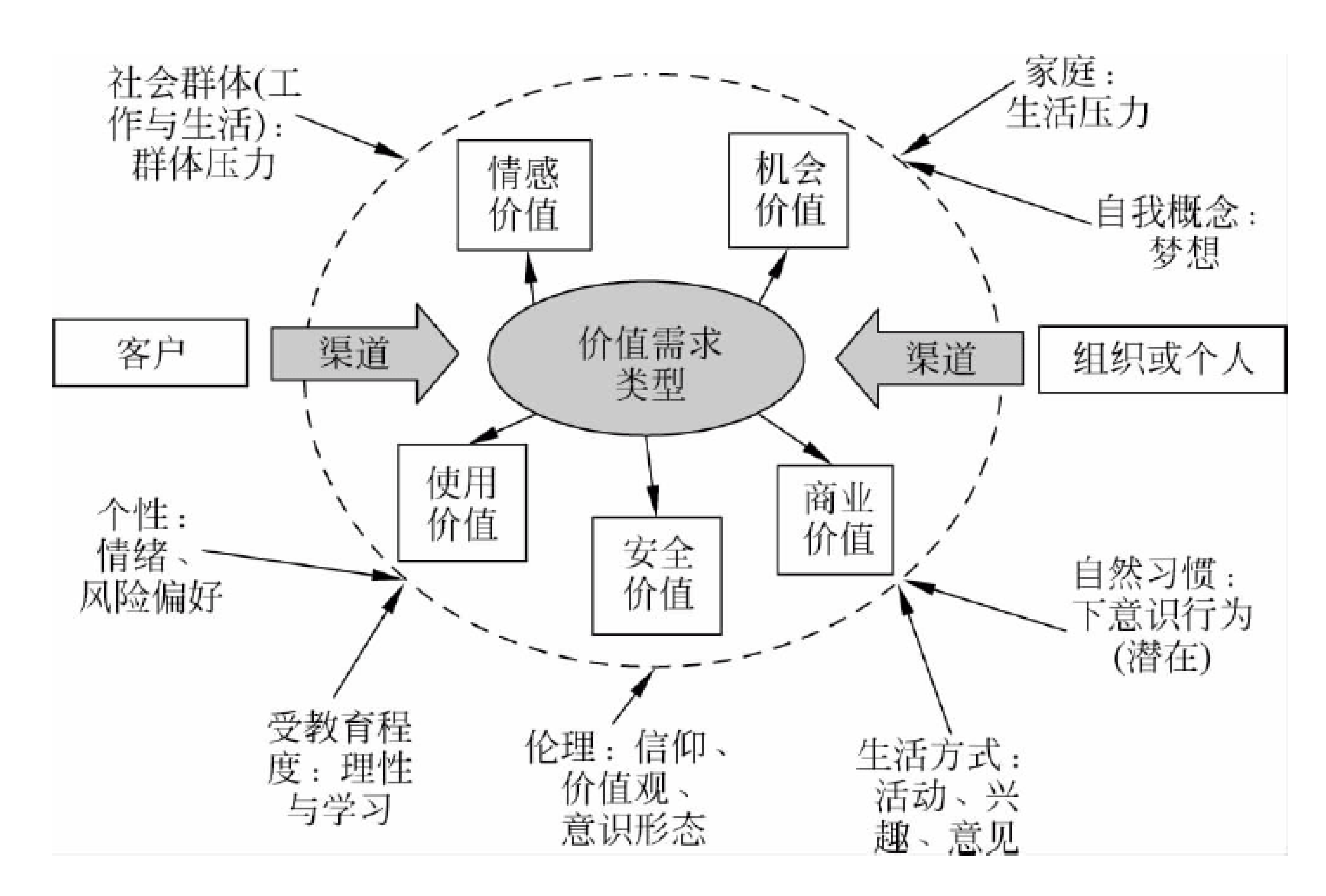


图 2-4　客户价值取向的决定因素模型

资金用于维护客户关系，过半数的品牌企业无法判别他们的最佳客户群。

(2) 客户分类管理。企业要根据需要对其拥有的客户进行合理的分类，并通过此分类建立起一对一的客户服务体系，实行差异化客户管理。客户分类的标准通常并不固定，可从定性和定量两个角度对客户进行分类。

① 定性的客户分类法。这是宏观上对企业所有的目标客户进行分类的一种方法。它是根据不同客户所认知的价值的侧重点不同对客户进行的分类。客户价值的形成一般可表示为：

$$\text{价值} = \text{利益} - \text{成本}$$

企业为了给客户提供更多的价值，就可以采用两种不同的方法，即提高利益或降低成本。那么，到底是为客户创造更多的利益好，还是提供价格更低廉的产品好，应该取决于客户的感觉。根据这种感觉的不同，可以把客户分为以下三类。一是内在价值型客户。这类客户的特点是对产品已有很深的了解，知道产品是否或在多大程度上满足他们的需求。他们只希望自己购买时所花费的费用合理，采购过程快捷便利，他们对各种建议和量身定做不感兴趣，低价格和便利的采购程序可以给他们带来最大价值和满足感。二是外在价值型客户。除了产品本身的价值外，这类客户更看重企业为他们提供的建议和个性化订制方案的价值。他们认为，销售人员的帮助和建议会为他们创造额外价值，并且也愿意为此支付额外费用。这类客户一般局限于大中客户身上，因为客户规模太小，创造的价值不足以弥补双方所付出的时间、金钱和精力。三是战略型价值客户。这类客户只可能限定在企业的少数几个最大的客户内。他们要求企业能为他们投入大量时间，并建立起战略伙伴联盟关系，这种联盟关系的长远利益是可观的。

② 定量的客户分类法。因为客户价值是客户管理中很重要的一个变量，我们可以利用这个变量对客户进行定量分类。由于影响客户价值的因素主要有三个，即客户生命周期、客户平均每次消费额和客户平均消费周期，为此，可以建立如下的数学模型：

$$\mathrm{CRV} = \frac{s}{t} \times T$$

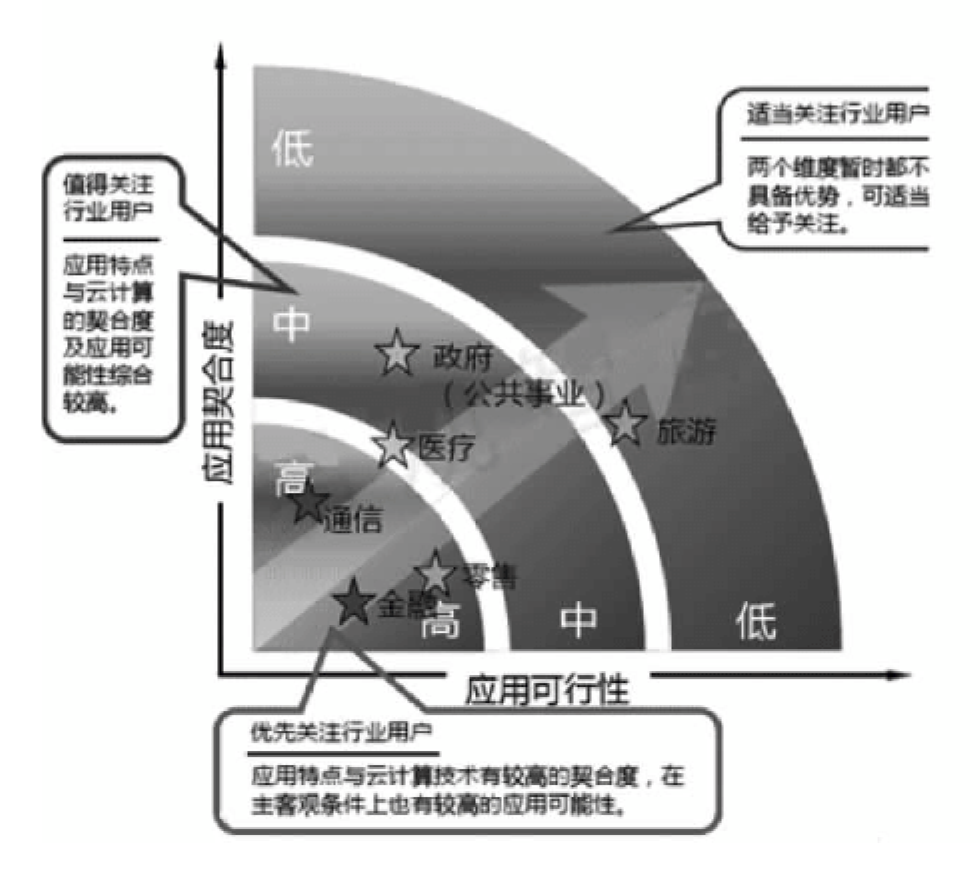
式中：CRV 为从核定期开始计算的客户生命周期的客户价值；T 为从核定期开始计算的客

户生命周期长度；s 为根据客户消费数据计算的客户平均每次消费额；t 为根据客户消费数据计算的客户平均消费周期。可见，客户价值主要取决于客户生命周期长度 T，客户平均消费周期 t 和客户平均每次消费金额 s，根据这三个指标的不同对客户进行如下分类。通常分为放弃客户、发展客户、白银客户和黄金客户。顾客中心论一直在追求每一个顾客的独特需求，实现一对一的服务，但受限于顾客影响因素太多和复杂，甚至连顾客自己也无法清晰定义需求；所以这一直是一个可望不可即的追求。不可否认的是技术进步一直是接近这一目标的驱动力，从 CRM、数据挖掘，到顾客在线定制，以至于顾客参与设计、参与创新，直接反馈产品和服务，不断推动着这一进程；今天，基于大数据，整合顾客消费、行为、生活数据，企业组织可以更好地提取顾客模式，提供个性产品和精准服务，提高顾客忠诚度；针对潜在顾客，可以进行精准营销，进行消费倾向管理，使其转化为企业真实客户；基于长尾效应，企业可以服务个性化小众市场，拓展市场空间。最重要的是顾客和市场是一个时刻变化的过程，每个人需求不尽相同，而且时刻在变化。每一个人在追求与众不同的同时，文化和社会结构又让大家寻求一致化，例如，一个强调客户个性化的企业是否要求员工统一服装？大数据无时无刻不在进行着市场分类、客户分类和管理对象的多维度分类；大数据使得组织得以考虑这种个性化、社群化及其动态变化共同决定的顾客、产品市场、员工市场、经理人市场、资本市场变化，并支持做出与之匹配的决策。大数据使得在企业组织在更多维度，根据不同场景和状态实施分类成为可能。

2.3 大数据市场的行业应用需求

根据中国大数据市场行业应用情况占比分析和应用成熟度分析，如图 2-5 所示，本节大数据行业应用情况主要结合热点行业和应用成熟度行业展开说明。





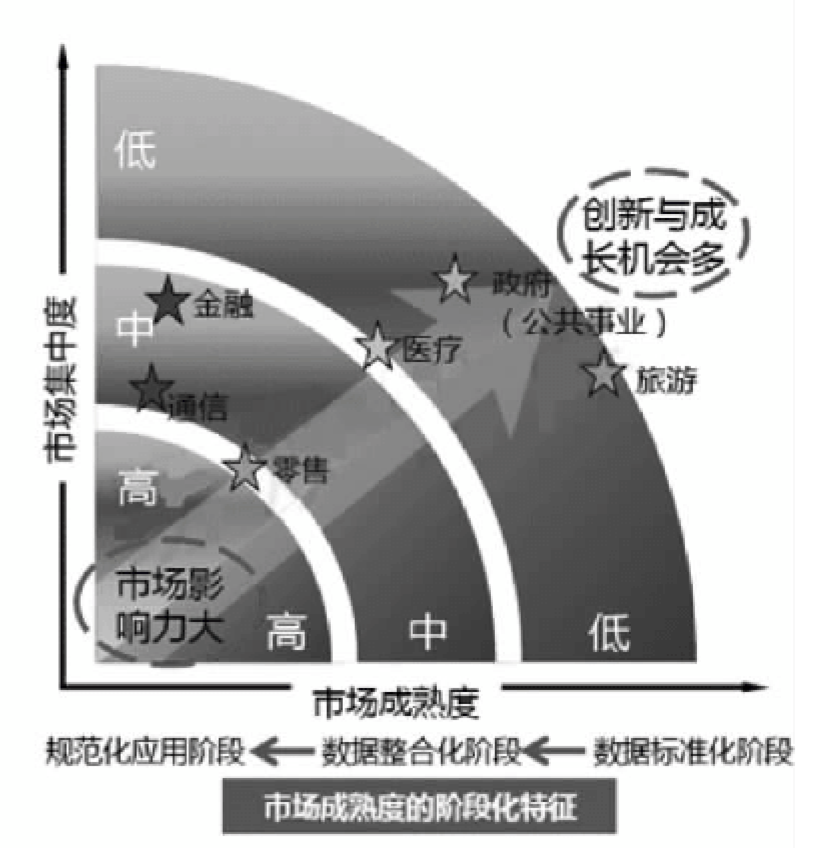


图 2-5　中国大数据市场行业应用情况分析

2.3.1 移动互联网和社交网络

在投身大数据的众多玩家中，电信运营商无疑是最为特殊的。庞大的网络规模和用户基础，提供了最全面的大数据样本；语音经营向流量经营的转型需求，以及来自互联网企业

和 OTT 业务冲击下的“被边缘化”危机，也使之具有推进大数据应用的迫切动力。

运营商的大数据可以分为三个层次，身份信息和账号等用户作为自然人个体的基础信息数据、用户使用运营商业务产生的行为信息数据，以及其他任何企业所不具备的网管日志和信号强度等基层网络数据。这些信息整合起来可以发现运营商所拥有的大数据具有非常典型的特性——既有用户真实生活的数据也有虚拟社会的实时数据，这是运营商大数据与互联网企业大数据的最大区别。在以全面性、实时性取胜的同时，运营商大数据的质量譬如关联性、可靠性也较高，而处理的单位成本更低，由此带来了更高的应用价值和挖掘潜力。

从运营商大数据的具体应用方向来看，当前主要集中在 4 个方向：流量经营精细化，智能客服中心建设，基于个性化服务的客户体验提升，以及对外数据服务等。

中国移动通信集团公司是中国规模最大的移动通信运营商，也是全球用户规模最大的移动运营商。基于大云平台构建了海量存储处理和数据分析和挖掘等核心能力，利用现有数据，探索大数据技术，已在河北等省试点，并尝试利用大数据技术识别异常话单。江苏移动建立了“智慧洞察”(Smart Insights)对外数据服务平台。该平台依托大数据强大的处理能力与海量数据，基于完全匿名和聚合后的数据，利用统计分析、数据挖掘等技术，提供标准化数据产品、大数据分析报告、高效 OpenAPI 服务。为社会、政府、企业以及家庭、个人客户提供经过分析挖掘而形成的价值产品与服务，实现数据价值提升与共享。

中国联通全面启动了以数据为中心的、集中化和一体化的 IT 系统建设，未来的建设模式转变为“平台＋应用”的模式，构建全集团唯一的、集中、开放的大数据平台，并在这个平台上构建各种各样的应用。把所有 IT 核心的数据、网元侧的数据、互联网的数据，乃至与外部合作和关联企业或者第三方交换的数据，全部整合，形成能够反映企业全景、客户全景、所有产品/渠道的大数据平台。这个平台采用大数据技术处理海量数据，并且能够将不同需求、不同业务有效整合，为上层应用提供定制化的服务。

大数据开启了电子商务行业的时代转型。电商和传统商家的最大区别在于：电商构建的各类型数据库能够涵盖商家信息、用户信息、行业资讯、产品使用体验、商品浏览记录、商品成交记录、产品价格动态等海量信息。电商行业大数据背后隐藏的是电商行业的用户需求、竞争情报，蕴藏着巨大的财富价值。借助大数据挖掘与分析技术，电商不仅可以提高营销转化为购买行为的成功率，还能降低营销成本，使产品更契合用户的需求，全面提升企业竞争力。

当前，我国主要的电子商务企业都在积极探索大数据应用，主要集中在以下方面。

(1) 大数据助推创新性平台化策略。消费群体的需求是多样并能被延伸的。为了服务这些延伸的消费需求，电商采用平台搭建方式，通过开放平台吸引第三方商家经营，能够提高电商平台渠道的利用效率，也丰富了电商平台的商品品类，满足用户延伸消费的需求，赢取竞争优势。以京东、阿里巴巴为首构建的全品类覆盖的综合性平台拉开了与其他中小型电子商务企业的差距。

(2) 市场预测。基于大数据预测技术能够实现产品从开发、生产、销售到物流整个链条的智能化和快速反应。通过对海量数据的收集、甄别与分析处理，不仅可以为终端的市场用户勾勒出消费习惯、消费能力的“用户画像”，大数据分析还能获取产品在各区域、各时间段、各消费群的库存和预售情况与发展趋势等，基于大数据预测技术能够实现从产品开发、生产、销售到物流等的整个链条的智能化和快速反应。

(3) 精准营销。大数据对电商行业的影响最典型的就是营销领域。对于电商来说，如何将产品定向推荐给需要的用户，始终是电商的核心关注点。消费数据量的急剧增加为电商企业精确把握用户群体和个体网络行为模式奠定了基础。电商企业通过大数据应用，划分细分群体与单体受众的心理层次，进行个人化、个性化、精确化和智能化广告推送与推广服务的探索。

(4) 用户体验。作为核心的服务理念，电子商务用户体验很大程度上决定了电子商务未来的成败。要让消费者最大限度地感受消费的归属感、满足感和幸福感，需要电商企业提供更智能、人性与差异化的服务，实现双赢的深度价值创造。在用户体验方面，各大电商包括垂直电商都形成了各具特色的个性化服务，用以提升用户体验。如基于用户画像为客户量身定做咨询应答策略，如快速理解用户意图、针对性商品评测或商品推荐、个性化关怀等。

(5) 物流与仓储优化。电子商务与物流业的合作随着云计算、物联网和数据应用等技术的突破越来越紧密，大数据改变了物流业的服务方向和服务内容，对于客户数据的分析也就不仅局限于电商企业单向操作。通过对客户数据的分析企业能够更合理地选择派送方式，优选路径，提供差异化服务，提高物流服务的质量，提升电商物流业的品牌形象。

2.3.2 政府公共管理

大数据的驱动力和引领作用正在给以政府为主导的公共管理领域带来革命性变化。在国家政策引导和支持下，各级政府和组织顺应大数据与云计算融合发展的技术和应用趋势，积极探索公共管理领域的大数据应用实践，大数据在支撑履行政府职能、保障公共安全、实施社会治理、支持重大决策和改进公共服务等方面发挥出越来越重要的作用。

1. 大数据有效支撑公共安全

公安行业经过“金盾工程”建设，形成了面向部门、警种的各类条线业务系统，以及面向多部门协作和底层信息支撑的综合业务系统，建成了可以全国范围内共享的 8 大业务信息资源库。公安行业积累的庞大数据，几乎和所有行业在数据层面都有密切交互。在大数据时代，走科技强警、信息化强警的大数据之路是解决警力不足等实际问题的重要途径。

为了应对大数据、云计算时代对下一代公安信息化建设的挑战，公安部多措并举，积极推进公安行业大数据应用实践。例如，在浙江警察学院建立了大数据应用重点实验室，2014 年公安部交通管理科学研究所与浪潮集团进行战略合作，建立交通管理大数据挖掘研判及云计算技术应用联合实验室。2012 年，山东省公安厅携手浪潮集团创新推出了大数据警务云计算中心建设工程，建成了“智能化全时空大数据预警系统”、新一代超级智能化搜索引擎“警务千度”等大数据应用，推动和引领了大数据和云计算技术融合的新一代公安行业信息化平台建设。

近年来，公安行业的大数据应用实践在提高反恐能力、预测犯罪趋势、推进案件侦破、破解交通难题等方面取得了比较丰硕的成果。例如，北京公安 110 指挥部积极探索践行大数据警务战略，自主研发了警情热点分布图等辅助指挥技术；上海交警利用大数据系统破解大城市的交通难题，圆满完成了 2014 年第 4 次亚信峰会的交通保障任务；苏州公安上线了依靠“大数据”理念建设了犯罪预测系统。总体来看，各级公安机关的大数据应用实践有效地提升了各警种的实战能力，大数据技术正在成为驱动和引领警务改革的关键要素。

2. 大数据改变工商管理模式

2015 年 4 月 28 日，基于大数据理念建设的国家工商总局广告数据中心正式启用，基本实现了对 31 个省(区市)332 个市所有媒体、全类广告的全覆盖、全天候监测，计划三年内监测范围将进一步扩展到全国两千八百多个县。各级工商管理部门广告监管机关，可依托该系统提供的大数据，实时掌握了解各地广告市场情况，及时派发监测发现的违法广告线索，形成证据提供、案件交办、立案查处、结果反馈一体化的监管指挥系统，提升广告监管执法效能。

此外，为探索大数据对提高市场主体的监管效率、规范市场秩序等方面的重要作用，国家工商总局先后与浪潮、百度、阿里、京东、龙信、拓普、海云、腾讯 8 家数据公司开展合作，并选定了 10 家试点单位先行，围绕主体监管和扶持小微企业发展两方面 11 项内容展开大数据分析应用。

目前，各试点单位积极采用大数据技术，通过内部集中工商组织数据、横向汇集其他政府部门数据、向外扩展关联互联网收集数据，构建新型监管模型和系统性风险防控机制，取得了显著成效。

3. 大数据帮助税源监控、税收预测、风险预警

在税务行业，当前已掌握了纳税人在税务登记、税务申报、税务稽查、税收评定等各个环节的海量数据信息，不仅如此，通过第三方数据交换渠道，税务机关还掌握了来自海关、工商、银行、统计、工信、公安、社保、财政等部门与纳税人生产经营有关的涉税数据。

基于掌握的海量数据，税务总局进行了以下积极探索。

(1) 税源监控。通过税源分析和挖掘，实时监控税收收入进度及税源变化情况，及时开展比对分析和检查评估，有效提升税源质量，减少税源流失，加强堵漏征收。

(2) 税收预测。通过宏观与微观数据分析相结合，加强税收政策、经济和税收关系分析，准确判断风险税收经济之间的关系，精准预测税收形势，科学估算税收收入规模，为组织收入工作提供依据。

(3) 风险预警。通过建立数据模型及信息综合比对，对纳税人生产经营活动中的涉税风险进行精准监控和提醒，形成以大数据为基础的风险识别、风险排序、风险分析、风险应对、绩效评价的完整闭环风险管理流程。

4. 大数据提供司法行业实证信息研究服务

在司法行业，人民法院以“大数据、大格局、大服务”理念为指导，运用顶层设计理念构建“数据集中管理平台”，建立数据全生命周期治理机制，整合全国法院司法信息资源；并运用语词提取、语义分析等现代化大数据分析技术，探索实证信息研究服务。

当前，已汇聚了全国 5000 万案件信息和 2400 万裁判文书信息，实现数据海量存储、科学分类、多元检索、深入分析，在第一时间提供涉众型经济犯罪、两抢一盗、强制医疗、知识产权纠纷等社会热点类案专项深度分析，探寻新形势下审判执行工作的特点和规律，促进社会治理创新。进一步发挥平台中枢作用，构建安全共享交换体系，提升各类数据在跨应用系统间、跨法院层级间、跨政府部门间、跨内外网系间的传输效率和共享水平，将更多的司法信息资源对社会公众公开，并为诉讼当事人和代理人提供司法大数据分析服务，加快构建开放、动态、透明、便民的阳光司法机制，让人民群众在每一个司法案件中切实感受到公平正义。

5. 用大数据保障公共安全

大数据的应用和发展可以帮助公共服务更好地优化模式，提升社会安全保障能力和面对突发情况的应急能力。作为大数据方面的开拓者——美国，在应用大数据来治理社会和稳定社会这方面的成绩显著。

美国国家安全局和交通安全局基于数据挖掘技术，开发了计算机辅助乘客筛选系统，为美国本土各个机场提供应用接口。该系统将乘客购买机票时提供的姓名、联系地址、电话号码、出生日期等信息输入商用数据库中，商用数据库则据此将隐含特殊危险等级的数字分值传送给交通安全局：绿色分值的乘客将接受正常筛选，黄色分值的乘客将接受额外筛选，红色分值的乘客将被禁止登机，且有可能受到法律强制性的关照。

同时，利用大数据也可预防犯罪案件的发生。加利福尼亚州桑塔克鲁兹市使用犯罪预测系统，对可能出现犯罪的重点区域、重要时段进行预测，并安排巡警巡逻。在所预测的犯罪事件中，有 2/3 真的发生。系统投入使用一年后，该市入室行窃率减少了 11%，偷车率减少了 8%，抓捕率上升了 56%。

另外，大数据也可以推进案件的侦破。这方面最经典的案例应该是波士顿连环爆炸案的成功告破。2013 年 4 月 15 日，美国波士顿在举办马拉松比赛的过程中发生连续炸弹爆炸案，导致 3 人死亡、183 人受伤。案件发生后警方不仅走访了事发地点附近 12 个街区的居民，收集可能存在的各种私人录像和照片，还大量收集网上信息，包括信息社交网站上出现的相关照片、录像等，并在这些网站上向公众提出收集相关信息的请求。通过对各方面数据的比对、查找，警方从录像中截取出了嫌疑人照片并发出通缉令，从而为最终追捕罪犯提供了确凿的证据和可靠的参考。

2.3.3 教育科研行业

教育大数据的主要目的是为不同利益相关者提供精准的教育服务，如学生的学习、教师的教学、开发者的资源开发、教育管理者的决策等；其核心是精准获取学习者的需求，为学习者提供精准教育服务；其数据主要来源于各类教育系统，包括学习管理系统（Learning Management System，LMS）、内容管理系统（Content Management System，CMS）、电子档案系统（e-Portfolio System，EPS）、智能培训系统（Intelligent Training System，ITS）、社会性学习系统（Social Learning System，SLS）、实时教学系统（Live Teaching System Based on Classroom，LTS）、学习设计系统（Learning Design System，LDS）和学生信息管理系统（Student Information System，SIS）等。应用和分析的教育大数据技术主要为教育数据挖掘技术和学习分析技术，当前研究热点为学习分析。学习分析是以理解和优化学习及其发生的环境为目的，对学习者及其所处情境的数据进行的测量、收集、分析和报告，其焦点是分析学习行为相关的数据、过程，以及呈现的方式。

国外主要机构相关标准制定与项目有美国"高级分布式学习"组织的 Experience API 标准、IMS Caliper Analytics 学习测评框架、欧盟学习分析项目 LACE（Learning Analytics Community Exchange，学习分析社区交流）等；国家标准组织 ISO/IEC JTC1SC36WG8 学习分析互操作工作组正在制定"学习分析互操作术语与参照模型"标准；而全国信息技术标准化技术委员会教育技术分委员会成立学习分析研究工作组开展教育大数据相关标准研究。

1. 利用大数据进行教育科研

科学数据是人类在认识自然、发展科技的活动中产生和积累的数据，是人类长期科学活动的知识积累，是一种重要的基础资源和战略资源。中国科学院作为中国自然科学的研究中心，在长期的科学研究实践中，通过观测、考察、实验、计算等多种途径产生和积累了大量具有重要科学价值和实用意义的科学数据和资料。

1983 年，中国科学院提出了“科学数据库及其信息系统”的建设项目，1986 年被国家计委列为国家“七五”和“八五”期间的重点工程项目。自此开始，经过中国科学院的持续支持，以及科技工作者的不懈努力，该项目取得了丰硕的成果，1997 年获中国科学院科技进步一等奖，1998 年获国家科技进步二等奖，到 2001 年“科学数据库及其信息系统”已经成为国内信息量最大、学科专业最广、服务层次最高、综合性最强的科学信息服务系统，成为科研工作的基础设施之一。“十一五”期间，中国科学院科学数据库的中国科学院信息化建设重要基础设施的定位进一步加固，各方面的工作都取得了重大进展与突破：数据资源建设了 51 个数据库，整合可共享的数据量达 148TB；服务环境基本形成了由网格运行服务总中心、学科领域网格主节点和数据资源网格节点三层架构的科学数据网格体系；基本完成了科学数据资源建设和服务标准体系的建设，同时研发部署了系列工具软件支撑标准规范的实施。科学数据库已初步形成结构合理的科学数据资源体系，并取得了数据资源整合服务的良好效果，以及社会应用效果。

“十二五”期间，该项目作为中国科学院信息专项“科技云”的重要内容，继续融合大数据和云计算等新技术建设“科技数据资源整合与共享工程”，截至 2014 年年底项目可共享总数据量超 450TB，并面向融合大数据和云计算等新技术继续完善标准规范体系。同时，项目继续延展服务科学研究和社会应用，重点支持了国家 863 计划、国家 973 计划、国家自然科学基金、国家科技支撑项目、国际合作项目、中国科学院创新性项目、先导专项等若干数据密集型的科研应用，起到了良好的示范作用，为科学研究、国家宏观决策、国民经济建设与社会发展等做出了重要贡献，产生了良好的社会效果。

大数据时代，科学模式已经变革为“数据密集型科学”的科研第四范式阶段，部分学科领域的科研活动已经成为典型的大数据行为，科学家有机会利用海量的科学数据去探索世界，开展此前无法进行的研究，解决此前难以解决的科学问题，产生突破性进展。近年来，国际上的一些科学研究成果充分证实了这一现实趋势，如生物领域和医学领域基于大规模 DNA 序列数据对生命现象的新认知，大型强子对撞机产生的海量实验数据帮助高能物理学家找得希格斯粒子等。科学领域曾是大数据的领先阵地，当前也正乘势快速发展中，未来科研大数据将是人类科研革命和社会进步的重要支撑。

2. 利用大数据促进教育行业变革

在教育工作中，特别是学校教育，数据成为教学改进显著的目标。美国国家教育统计中心已经把中小学和大学的学生学习行为、考试分数和职业规划等重要的数据存储起来，用于统计和分析。而近年来越来越多的网络在线教育和大规模开放式网络课程的兴起，使教育领域中的大数据获得了更为广阔的应用空间。

教育领域中大数据分析的最终目的是提高学生的学习成绩。美国教育部门创造了一套“学习分析系统”，将教育和大数据相结合。该系统是一个数据挖掘和案例运用的联合框架，

主要向教育工作者提供影响学习成绩的原因等信息，为教师提供提高学生成绩更准确有效的办法。

美国已经存在一些企业成功地商业化运作了教育中的大数据。例如，IBM 与亚拉巴马州的莫白儿县公共学区在大数据方面展开合作，从而较好地改善了该学区的辍学情况；希维塔斯学习(Civitsa Learning)在高等教育领域建立了最大的跨校学习数据库，通过这些海量数据，可以看到学生的分数、出勤率、辍学率和保留率等数据的主要趋势；梦盒学习(Dream Box Learning)公司和纽顿(Knewton)公司已经成功创造并发布了各自的利用大数据的适应性学习系统。

在我国，百度推出了"百度预测"，在 2014 年也通过数据分析，预测出高考作文题目的出题范围将会在"生命的多彩""时间的馈赠"等 6 个领域中，并且给出了各领域命中的精确概率。对试题的精确预测，也可以较大程度上提高学生的学习成绩。

2.3.4 金融行业

面对互联网金融的竞争压力，金融企业急需重构以金融大数据分析为基础的决策和服务体系，提升自身竞争力和客户满意度。在大数据时代，银行数据量不断增加，现有以交易为核心的数据处理系统，无法满足大数据处理的要求。金融企业更需要建设第二数据平面，以处理更多维、更大量的数据。

金融大数据典型应用场景包括：历史交易明细查询、实时征信、实时事件营销、客户行为分析等。大数据解决方案围绕金融大数据的采集、存储、处理、洞察和服务，为银行开发新业务，提供业务支撑，激发金融创新活力。解决方案提供的主要功能包括：①海量结构化/非结构化数据的采集、存储、批处理、内存计算和实时流计算的能力；②百万维大数据特征提取、管理、建模的能力，帮助客户直接实现小微贷款预测或金融资产预测等业务；③历史交易明细查询、实时征信、实时事件营销等，让客户更专注大数据业务开发本身，更方便地使用大数据能力。并具有易用性和复杂查询能力，可实现与银行现有数据库、数据仓库的无缝对接。此外，需根据银行生产系统的规范，在大数据的可靠性、安全性、易用性方面进行了增强和适配，如支持金融数据异地容灾等。

数据分析在金融业中最直接的应用是个人信用等级的评估。美国个人消费信用评估公司 FICO 在 20 世纪 50 年代发明了信用积分概念和评价方法，根据支付历史、欠款金额和使用信用卡时间长度等信用报告指标进行信用评分，并用于个人信贷等领域。进入大数据时代后，越来越多的新指标被纳入评估体系，包括过去常常被认为是不可能获取的社交网络数据。

银行、证券和保险是金融类企业的三个重要部分。国内不少银行已经开始尝试通过大数据来驱动业务运营。例如民生银行，其 80%以上的客户是小微企业。借助大数据平台，民生银行的每家小微企业客户的信息都能够实时上报民生的"数据加工厂"，并生产出有价值的信息，使总行能够更加快速、准确地获得各个行业的市场需求信息，从而快速、精确地进行战略决策和市场规划。

基于大数据平台，民生银行实现了内部管理的精细化，"用数据说话、靠数据决策"已经成为民生银行的一种管理文化。依据大数据平台和专业金融技术工具，民生银行目前能够准确计算出每位客户的利润贡献度，从而真正做到个性化定价和个性化服务。在产品定价

方面，以往银行都按照批量定价模式，向客户销售贷款；而个性化定价，则根据客户的存款、贷款、业务经营情况等综合指标进行科学定价，不仅能够吸引优质客户，提高客户黏性，降低客户流失率，也能够提高整体收益。基于大数据平台，民生银行实现了从“广撒网”到“批量定向开发”的转变。除了民生银行，光大银行建立了社交网络信息数据库，招商银行利用大数据发展小微贷款，中信银行信用卡中心使用大数据技术实现了实时营销。

在证券行业，大数据主要包含几个方面的应用：股价预测、客户关系管理和投资景气指数。

现在很多股权的交易都利用大数据算法进行，这些算法现在越来越多地考虑了社交媒体和网站新闻来决定在未来几秒内是买入还是卖出。IBM 日本的新系统仅用 6 小时就预测出分析师需要花费数日才能计算出的预测值，它结合其他相关经济数据的历史数据分析与股价的关系，从而得出预测结果。

对客户关系的管理包括两个方面，对客户进行细分和客户流失的预测。通过对客户的账户状态进行分析，对客户进行聚类和细分，从而发现客户交易，找出最有价值和盈利潜力的客户群，为他们提供个性化服务。证券公司通过对客户的历史交易行为和流失情况进行分析，建立客户流失模型，从而预测客户流失。

在保险行业，大数据应用也包括三个方面：客户细分及精细化营销、欺诈行为分析和精细化运营。例如，友邦保险使用了大数据魔镜软件，开发出客户挖掘、精准投放、二次开发、战略指导、全民分析等多种智能分析模型，为管理层提供最直接的数据依据，之前每个保险业务员从 200 个电话中，可能才能挖掘出两三个意向客户，而精准的投放使得平均拨打一个电话就可以得到一个客户。

2.3.5 医疗健康业

人体是十分复杂的系统。传统医学尤其是西医，注重了解人体的内部构成，研究疾病成因并施以治疗。而在海量数据的帮助下，相关关系的挖掘变得更加简单、快捷、准确，在采集海量数据的情形下，医生甚至可以直接依据相关关系进行疾病的预判和诊疗。

在 IBM、安大略理工大学和一些医疗的合作项目中，心率、呼吸、体温、血压和血氧含量等 16 组数据被用于检查早产儿的身体状况，这些数据的采集频度达到了每秒 1260 次之多，在这个系统的帮助下，医生可以通过早产儿的身体细微变化预判他们可能出现的感染症状，将诊断预防提前 24 小时。

而与此同时，IBM 也和其他机构就大数据应用开展了多项合作。在与美国加州大学洛杉矶分校里根医学中心的医生们就创伤性脑损伤治疗的合作中，IBM 的科学家通过分析从患者身上获得的巨大数据流，预测出现可能导致认知能力损害至死亡的脑肿胀病情的可能性。通过跟踪实时采集到的患者呼吸率数据和心率模式，医生可以利用 IBM 开发的大数据软件识别并预测患者未来数小时的各种生理迹象。

IBM 大数据项目负责人 NaguiHalim 说：“我可以将治疗脑损伤的大数据技术与一本书的写作做一个生动的比较。计算机科学家通常会在数据被编译后才对其进行分析——就像扫描已完成的一本或者一百本书的关键字一样。有了目前的技术，我们可以一边打字，一边分析。”Halim 还说，未来科学家可以通过研究病人的病历，将病人的健康形态拼在一起来预测病人未来的状况——就像可以通过了解一个作者如何在其以往著作中塑造任务和故事

情节，从而在其未成书时预测书的内容。得克萨斯州的脑损伤专家 BrentMasel 医生表示："这并不能彻底治愈脑损伤，但是它在脑损伤恢复上却起到了非常了不起的作用。这使我们的治疗更为精确，意义非凡。"

哈佛医学院布赖海姆女子医院的医学研究人员也在使用大数据技术来研究开给 1000 万患者处方药的效果。研究人员正在创建全新的研究方式来分析海量数据，用以辨别数以百万计病患者的用药风险。

对数据分析能力的增强也使得更精细的诊断分析成为可能，在苹果公司前总裁史蒂夫·乔布斯的癌症治疗过程中，他支付几十万美元的费用完成了自身所有 DNA 序列和肿瘤序列的排序，以便医生们能够基于他的个体基因组成给出用药建议。

谷歌的 FluTrend 可以利用搜索关键词和大数据技术成功预测流感的散布趋势。在流感爆发前，人们用谷歌搜索流感的相关资讯或措施的比例将会增加，谷歌通过对无数流感关键词进行分析，可以准确快速地预测流感将在哪里出现，以及流感的散布范围。这一项目的成功也刮起了大数据变革公共卫生的浪潮。目前，谷歌又孵化了一个医疗健康项目，名为 Baseline，它主要用大数据来预防癌症。

百度公司也在疾病预测方面做了一些工作。2014 年 7 月，在百度推出世界杯预测之后，又上线了一个最新服务：疾病预测。它能为用户提供流感、肝炎、肺结核和性病 4 种疾病的趋势预测，并可根据过去 30 天的资料，对未来 7 天疾病变化进行预测。目前该服务已经涵盖了中国 331 个城市，2870 个区县，并且某些城市已经细化到以商圈为目标单位，未来甚至可以细化到个人的粒度。

对于目前正在爆发的埃博拉病毒，也可以通过大数据技术来预防疾病的传播，对疫情进行更好的控制，做好民众的救助工作。首先，西非等地的跨国电信业者与国际卫生组织合作，提供当地居民行为通信资料，通过分析绘制当地居民聚落位置和人口移动地图，来预测病毒散布的位置。其次，非洲政府可以根据用户的手机定位，分析出当地居住区位置的移动轨迹，规划医疗救助站的位置，从而安排最佳的救助路线，使居民远离疫情较为严重的区域。

除了在疾病预测方面，利用大数据的计算和分析能力，能够让我们在几分钟内解码整个 DNA，制定出最新的治疗方案。大数据技术目前已经在医院应用监视早产婴儿和患病婴儿的情况，通过记录和分析婴儿的心跳，医生针对婴儿的身体可能会出现的不适症状做出预测，这样可以帮助医生更好地救助婴儿。

大数据已经在医疗和健康领域取得了一定的成果，将疾病防治关口前移，可以大大节省医疗资源的消耗。有效的数据分析也可以提前对民众进行医疗健康知识的普及教育，从而较好地预防疾病的发生。

2.3.6 中国制造 2025

制造业是国民经济的主体，是立国之本、兴国之器、强国之基。18 世纪中叶开启工业文明以来，世界强国的兴衰史和中华民族的奋斗史一再证明，没有强大的制造业，就没有国家和民族的强盛。打造具有国际竞争力的制造业，是我国提升综合国力、保障国家安全、建设世界强国的必由之路。

新中国成立尤其是改革开放以来，我国制造业持续快速发展，建成了门类齐全、独立完整的产业体系，有力推动工业化和现代化进程，显著增强综合国力，支撑我世界大国地位。

然而，与世界先进水平相比，我国制造业仍然大而不强，在自主创新能力、资源利用效率、产业结构水平、信息化程度、质量效益等方面差距明显，转型升级和跨越发展的任务紧迫而艰巨。

当前，新一轮科技革命和产业变革与我国加快转变经济发展方式形成历史性交汇，国际产业分工格局正在重塑。必须紧紧抓住这一重大历史机遇，按照“四个全面”战略布局要求，实施制造强国战略，加强统筹规划和前瞻部署，力争通过三十年的努力，到新中国成立一百年时，把我国建设成为引领世界制造业发展的制造强国，为实现中华民族伟大复兴的中国梦打下坚实基础。

加快推动新一代信息技术与制造技术融合发展，把智能制造作为两化深度融合的主攻方向；着力发展智能装备和智能产品，推进生产过程智能化，培育新型生产方式，全面提升企业研发、生产、管理和服务的智能化水平。

研究制定智能制造发展战略。编制智能制造发展规划，明确发展目标、重点任务和重大布局。加快制定智能制造技术标准，建立完善智能制造和两化融合管理标准体系。强化应用牵引，建立智能制造产业联盟，协同推动智能装备和产品研发、系统集成创新与产业化。促进工业互联网、云计算、大数据在企业研发设计、生产制造、经营管理、销售服务等全流程和全产业链的综合集成应用。加强智能制造工业控制系统网络安全保障能力建设，健全综合保障体系。

加快发展智能制造装备和产品。组织研发具有深度感知、智慧决策、自动执行功能的高档数控机床、工业机器人、增材制造装备等智能制造装备以及智能化生产线，突破新型传感器、智能测量仪表、工业控制系统、伺服电机及驱动器和减速器等智能核心装置，推进工程化和产业化。加快机械、航空、船舶、汽车、轻工、纺织、食品、电子等行业生产设备的智能化改造，提高精准制造、敏捷制造能力。统筹布局和推动智能交通工具、智能工程机械、服务机器人、智能家电、智能照明电器、可穿戴设备等产品研发和产业化。

推进制造过程智能化。在重点领域试点建设智能工厂/数字化车间，加快人机智能交互、工业机器人、智能物流管理、增材制造等技术和装备在生产过程中的应用，促进制造工艺的仿真优化、数字化控制、状态信息实时监测和自适应控制。加快产品全生命周期管理、客户关系管理、供应链管理系统的推广应用，促进集团管控、设计与制造、产供销一体、业务和财务衔接等关键环节集成，实现智能管控。加快民用爆炸物品、危险化学品、食品、印染、稀土、农药等重点行业智能检测监管体系建设，提高智能化水平。

深化互联网在制造领域的应用。制定互联网与制造业融合发展的路线图，明确发展方向、目标和路径。发展基于互联网的个性化定制、众包设计、云制造等新型制造模式，推动形成基于消费需求动态感知的研发、制造和产业组织方式。建立优势互补、合作共赢的开放型产业生态体系。加快开展物联网技术研发和应用示范，培育智能监测、远程诊断管理、全产业链追溯等工业互联网新应用。实施工业云及工业大数据创新应用试点，建设一批高质量的工业云服务和工业大数据平台，推动软件与服务、设计与制造资源、关键技术与标准的开放共享。

加强互联网基础设施建设。加强工业互联网基础设施建设规划与布局，建设低时延、高可靠、广覆盖的工业互联网。加快制造业集聚区光纤网、移动通信网和无线局域网的部署和建设，实现信息网络宽带升级，提高企业宽带接入能力。针对信息物理系统网络研发及应用

需求，组织开发智能控制系统、工业应用软件、故障诊断软件和相关工具、传感和通信系统协议，实现人、设备与产品的实时连通、精确识别、有效交互与智能控制。

福特公司内部每一个职能部门都会配备专门的数据分析小组，同时还在硅谷设立了一个专门依据数据进行科技创新的实验室。这个实验室收集大约四百万辆装有车载传感设备的汽车数据，通过对数据进行分析，工程师可以了解司机在驾驶汽车时的感受、外部的环境变化以及汽车环境的相应表现，从而改善车辆的操作性，提高能源的利用效率和车辆的排气质量，同时，还针对车内噪声的问题改变了扬声器的位置，从而最大程度减少了车内噪声。在 2014 年举行的北美国际车展中，福特重新设计了 F-150 皮卡车，使用轻量铝代替了原来的钢材，有效减少了燃料消耗。负责 F-150 皮卡车设计的数据分析师 Michael Cavaretta 说，在减少燃料消耗的过程中，技术团队选择了多项备选方案，并估算了这些技术的成本和利润，以及实现技术需要消耗的时间的基础上进行了优化分析和抉择，而轻量铝就是团队在进行了数据分析和综合评估之后的选择。

福特研究和创新中心一直希望能够通过使用先进的数学模型帮助福特汽车降低对环境的影响，从而提高公司的影响力。针对燃油经济性问题，这个由科学家、数学家和建模专家所组成的研究团队开发出了基于统计数据的研发模型，对未来 50 年内全球汽车所产生的二氧化碳排放量进行了预测，进而帮助福特公司制定较高的燃油经济性目标并提醒公司高层保持对环境的重视。针对汽车能源动力选择问题，福特数据团队利用数学建模方法，证明某一种替代能源动力要取代其他多有动力的可能性很小，由此帮助福特开发出包括 EcoBoost 发动机、混合动力、插电式混合动力、灵活燃料、纯电动、生物燃油、天然气和液化天然气在内的一系列动力技术。同时福特团队还开发了具有特殊功能的分析工具，如福特车辆采购计划工具，该分析系统能根据大宗客户的需求帮助他们进行采购分析，同时也帮助他们降低成本和保护环境。福特认为分析模型和大数据将是增强自身创新能力、竞争能力和工作效率的下一个突破点，在越来越多新的技术方法不断涌现的今天，分析模型与大数据将为消费者和企业自身创造更多的价值。

2.3.7 智能交通领域

随着大数据时代的到来，智能交通迎来重大变化，智能交通产业发展也将迎来新的机遇。交通拥堵、交通污染日益严重，交通事故频繁发生，这些都是各大城市亟待解决的问题，智能交通成为改善城市交通的关键所在。及时、准确地获取交通数据并构建交通数据处理模型是建设智能交通的前提，而这一难题可以通过大数据技术得到解决。

法国里昂市与 IBM 的研究者合作开发出能够缓解道路拥堵的系统方案。IBM 为里昂开发的系统名为 Decision Support System Optimizer(决策支持系统优化器)，可以基于实时的交通情况报告来侦测和预测交通拥堵。当交管人员发现某地即将发生交通拥堵时，可以及时调整信号灯让车流以最高效率运行。这个系统对于突发事件也很有用，例如帮助救护车尽快到达医院。而且随着运行时间的积累，这套系统还能够“学习”过去的成功处置方案，并运用到未来预测中。

SpotHero 是预订停车位的一个移动应用，它的网站和移动应用可以较好地解决司机找不到停车位的问题。SpotHero 能够实时跟踪停车位数据变化，打开 SpotHero，将会显示附近可用的停车位的公交车和价格，同时提供导航服务，并且可以使用预付费来占领未被使用

的停车位。目前,已经能够实时监控包括华盛顿、纽约、芝加哥、巴尔的摩、波士顿、密尔沃基和纽瓦克 7 个城市的停车位。

大数据在智能交通应用方面的优势体现在:大数据技术的海量数据存储和高效需求分析能力,能够实现交通管理系统跨区域、跨部门的集成和组合,更有效地配置交通资源。大数据的实时性,使处于静态闲置的数据被处理和需要利用时,即可被智能化利用,使交通运行得更加合理,从而提升交通运行效率和服务的水平。其次,大数据技术具有较高的预测能力,可降低误报和漏报的概率,随时针对交通的动态性给予实时监控。基于对大数据的预测性分析,通过梳理影响安全运行的各种原因,发现道路运行安全管理的内在规律,将为交通管理决策、规划、运营、服务,以及主动安全防范带来更加有效的支持,以提高交通安全的水平,在一定程度上避免交通事故。此外,大数据技术在减轻道路交通堵塞、降低汽车运输对环境的影响等方面有重要的作用。通过建立区域交通排放的监测及预测模型,共享交通运行与环境数据,建立交通运行与环境数据共享实验系统,大数据技术可有效分析交通对环境的影响。同时,通过分析历史数据,大数据技术能提供降低交通延误和减少排放的交通信号智能化控制的决策依据,建立低排放交通信号控制原型系统与车辆排放环境影响仿真系统。

面对海量的交通信息,交通大数据的开发应用需求日益突出,交通大数据时代的来临是智能交通发展的必然趋势,这将为智能交通提供更多的发展机遇和空间。



大数据平台的架构体系

一个完整的大数据平台，其架构体系一般由如图 3-1 所示的几部分组成。

安全管理	大数据应用	运营管理
	数据交互展示	
	数据处理	
	数据存储	
	数据采集	
	大数据基础设施	

图 3-1　大数据平台架构体系

其中，大数据基础设施是大数据存储、计算、展示等的基础支撑设施；数据采集是把数据从数据源采集导入到数据平台中的相关接口及技术；数据存储则是将数据采用分布式文件、分布式数据库的方式存储在大规模的节点中；数据处理是对所存储的数据进行查询、统计、分析、预测、挖掘、商业智能处理、深度学习等相关处理；数据交互展示则是将分析处理完的数据以最佳的交互方式呈现给数据使用者和消费者；大数据应用是把数据及处理结果应用到各行各业中去，比如医疗、环保、社交、金融、中国制造等行业；安全管理是对数据的全方位安全管控；运营管理则是保障整个数据处理架构的稳定高效运营。

下面的章节中将逐一详细介绍相关的组成部分。

3.1　大数据基础设施

大数据基础设施为大数据平台的底层提供必要的基础设施支持，比如基础的计算、存储、网络设备，云数据中心，云计算平台等。基础设施与大数据处理的关系，就像我们的身体为大脑思考提供能量一样。强健的体魄可以为大脑提供充足的能量支持，而完善的基础设施可以支持强大的数据处理。

大数据处理需要拥有大规模物理资源的云数据中心和具备高效的调度管理功能的云计算平台的支撑。云计算管理平台能为大型数据中心及政府、企业提供灵活高效的部署、运行和管理环境，通过虚拟化技术支持异构的底层硬件及操作系统，为应用提供安全、高性能、高可扩展、高可靠和高伸缩性的云资源管理解决方案，降低应用系统开发、部署、运行和维护的

成本，提高资源使用效率。

3.1.1 虚拟化

虚拟化是在 1960 年为了描述虚拟机（实验性的 IBM M44/44X 系统）这个概念时被第一次提出的。虚拟化的概念也比较好理解，在电影《黑客帝国》中，男主角尼奥（Neo）生活在由一台超级计算机母体（Matrix）所创造出来的模拟世界中，在里面上班工作，后来逐步醒悟到他只是活在机器所设定的一个虚拟世界里，最终率领人类摆脱机器的控制。这里面的虚拟世界就是对现实世界的一种模拟，在里面所有的体验都跟在真实世界中的一样。

按虚拟化技术的应用特点，虚拟化技术主要分为以下几类：服务器虚拟化、存储虚拟化、网络虚拟化及桌面虚拟化。将虚拟化技术应用于数据中心领域，能够解决阻碍数据中心发展的诸多问题，提高物理设备的利用率，有效降低数据中心运维成本，降低能耗以及保证数据中心服务的可靠性、连续性。

对虚拟机的构建和管理被称为平台虚拟化，现在也称为服务器虚拟化。平台虚拟化，跟上面的虚拟世界类似，就是在一个给定硬件平台的服务器（宿主机）上创造一个模拟的计算机环境（虚拟机），并提供给客户机。许多宿主机允许运行真实的操作系统，客户机就好像直接运行在宿主机的计算机硬件上，而实际上它是运行在虚拟机上。一般虚拟机对硬件资源（如网络、显示器、键盘、硬盘）的访问被统一管理在一个比处理器和系统内存更有限制性的层次上。客户软件经常被限制访问计算机周边设备，或者被限制在较低的设备性能上，这取决于宿主机硬件访问策略设定。

采用虚拟化技术有几个方面的原因。一方面根据摩尔定律和 CPU 生产技术的迅猛发展，当今的计算机性能越来越强大，配置越来越高。比如市场上常见的一款智能手机往往都是 4 核甚至 8 核的 CPU，其计算和存储能力远超一台最早期的超级计算机。但这样强大的硬件和处理能力仍然被一个统一的操作系统管理，造成资源和效率的浪费。为了发挥所有的 CPU 和硬件资源的效率，可以把每一个运行在独立的服务器上的操作系统转移到虚拟机中。大型的服务器可以“寄宿”许多这样的“客户”虚拟机。这就是物理到虚拟（Physical-to-Virtual，P2V）的转换。

另一方面虚拟机相比于物理机器，具备很多的优势和灵活性。比如虚拟机可以被更容易地从外部被控制和检查，并且可以更灵活地配置（CPU 核数、内存、硬盘、网络等）和升级维护。

另外，创建一个新的虚拟机不需要预先购买硬件。同时，一个新的虚拟机可以容易地从一台计算机转移到另一台上。一个销售员可以很方便地把一个包含试用版软件的虚拟机复制到他的笔记本中，再去拜访他的客户时不用更换计算机。类似地，虚拟机中的故障不会对宿主机产生损害，所以不会令笔记本上的操作系统死机。

虚拟机由于可以很容易地迁移，所以也常被用于远距离灾难恢复方案。

3.1.2 云计算

云计算是继 20 世纪 90 年代大型计算机到客户端-服务器的大转变之后的又一种巨变。由于政府和企业用户对于大型计算资源的需求在不断上升，而他们自己独立购买、建设和运

营大规模的服务集群的成本又非常高昂，因而诞生了大型的第三方云数据中心服务商，为用户提供云计算服务。云计算基于的经济模式是规模经济效应，也就是说很多的中小用户在云资源平台上共享资源，这样云服务商可以综合盈利。这种商业模式类似于现今的电网和自来水管网。在电网系统中，有大型的发电厂，通过输变电路把电接入企业和千家万户，我们只需按用电量来支付电费。在自来水供应中类似地有大型自来水厂，通过输送管网，传送到用户家中，我们打开水龙头就能用水，按照用水量支付水费。

云计算的模式也类似，用户接入网络，就能使用大型云数据中心里的存储和计算资源，而不再需要了解“云”中基础设施的细节，不必具有相应的专业云计算知识，也无须直接进行控制。云计算描述了一种基于互联网的新的 IT 服务增加、使用和交付模式，通常涉及通过互联网来提供动态、易扩展而且经常是虚拟化的服务。

随着信息和通信技术的快速发展，如图 3-2 所示，计算模式经历了从最初把任务集中交付给大型处理机模式，到后来发展为基于网络的分布式任务处理模式，再到最新的按需处理的云计算模式。最初的单个处理机模式处理能力有限，并且请求需要等待，效率低下。后来，随着网络技术的不断发展，按照高负载配置的服务器集群，在遇到低负载的时候，会有资源的浪费和闲置，导致用户的运行维护成本提高。而云计算把网络上的服务资源虚拟化，整个服务资源的调度、管理、维护等工作由专门的人员负责，用户不必关心“云”内部的实现，因此云计算实质上是给用户提供像传统的电力、水、煤气一样的按需计算服务，它是一种新的有效的计算使用范式。并且，云计算是分布式计算、效用计算、虚拟化技术、Web 服务、网格计算等技术的融合和发展，其目标是用户通过网络能够在任何时间、任何地点最大限度地使用虚拟资源池，处理大规模计算问题。目前，在学术界和工业界的共同推动之下，云计算及其应用呈现迅速增长的趋势，各大云计算厂商如 Amazon、IBM、Google、Microsoft、Sun 等公司都推出自己研发的云计算服务平台。而学术界也源于云计算的现实背景纷纷对模型、应用、成本、仿真、性能优化、测试等诸多问题进行了深入研究，提出了各自的理论方法和技术成果，极大地推动了云计算继续向前发展。

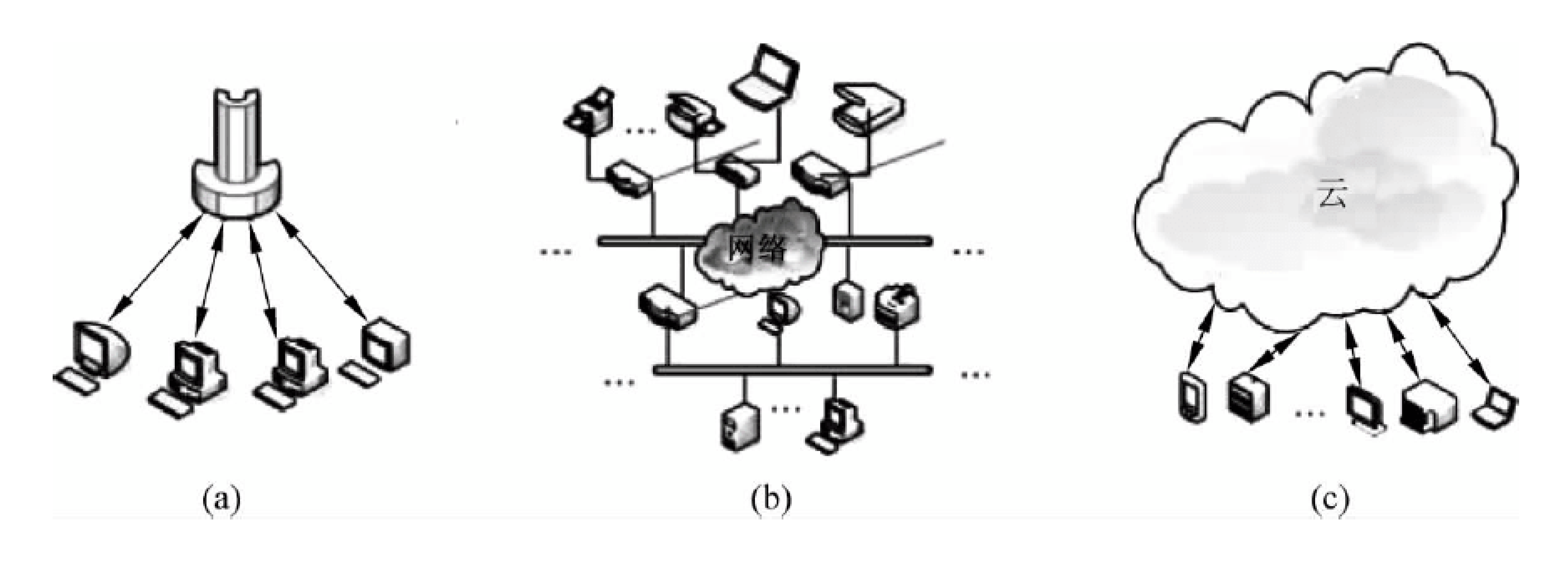


图 3-2　云计算模式的演化

1. 云计算定义

云计算概念最早是由 Google 提出的，一方面是因为当时在网络拓扑图中用云来代表远程的大型网络，另一方面用来指代通过网络应用模式来获取服务。狭义云计算是指 IT 基础设施的交付和使用模式，指通过网络以按需、易扩展的方式获得所需的资源；广义云计算是指服务的交付和使用模式，指通过网络以按需、易扩展的方式获得所需的服务。这种服务

可以是 IT 和软件、互联网相关的，也可以是任意其他的服务，它具有超大规模、虚拟化、可靠安全等独特功效。

目前，不同文献和资料对云计算的定义有不同的表述，主要有以下几种代表性的定义。

定义 1　云计算是一种能够在短时间内迅速按需提供资源的服务，可以避免资源过度和过低使用。

定义 2　云计算是一种并行的、分布式的系统，由虚拟化的计算资源构成，能够根据服务提供者和用户事先商定好的服务等级协议动态地提供服务。

定义 3　云计算是一种可以调用的虚拟化的资源池，这些资源池可以根据负载动态重新配置，以达到最优化使用的目的。用户和服务提供商事先约定服务等级协议，用户以用时付费模式使用服务。

定义 4　云计算是一种大规模分布式的计算模式，由规模经济所驱动，能够把抽象化的、虚拟化的、动态可扩展的计算、存储、平台及服务以资源池的方式管理，并通过互联网按需提供给用户。

定义 1 强调了按需使用方式，定义 2 中突出了用户和服务提供商双方事先商定的服务等级协议。这两个定义都从一定的角度给出定义。定义 3 和定义 4 综合了前面两种定义的描述，更好地揭示了云计算的特点和本质。

2. 云计算主要特征

云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池(资源包括网络、服务器、存储、应用软件、服务)，这些资源能够被快速提供，只需要投入很少的管理工作，或与服务供应商进行很少的交互。云计算有以下 5 个主要特征。

(1) 按需自助服务。消费者可以单方面按需部署处理能力，如服务器时间和网络存储，而不需要与每个服务供应商进行人工交互。

(2) 通过网络访问。可以通过互联网获取各种能力，并可以通过标准方式访问，以通过众多瘦客户端或富客户端推广使用(例如移动电话、笔记本、PDA 等)。

(3) 与地点无关的资源池。供应商的计算资源被集中，以便以多用户租用模式服务所有客户，同时不同的物理和虚拟资源可根据客户需求动态分配和重新分配。客户一般无法控制或知道资源的确切位置。这些资源包括存储、处理器、内存、网络带宽和虚拟机器。

(4) 快速伸缩性。可以迅速、弹性地提供资源，能快速扩展，也可快速释放以实现快速缩小。对客户来说，可以租用的资源看起来似乎是无限的，并且可在任何时间购买任何数量的资源。

(5) 按使用付费。能力的收费是基于计量的一次一付，或基于广告的收费模式，以促进资源的优化利用。比如计量存储，带宽和计算资源的消耗，按月根据用户实际使用收费。在一个组织内的云可以在部门之间计算费用，但不一定使用真实货币。

云计算新的范式的特点带来了众多的优势，同时引入了一些新的问题亟待解决。这些因素制约着云计算技术及其应用的发展，见表 3-1。

表 3-1　云计算的优势和对应问题

云计算	优　　势	问　　题
安全性	缩短单机密集数据处理任务时间，把处理任务分配到各个节点计算，提高了效率	用户关注传输到云计算端的敏感处理数据是否安全
可靠性	减少用户购买物理硬件设备的费用，资源以服务的方式进行租赁，降低用户资金投入的前期风险，促进用户把精力投入业务中	虽然用户不需要维护软件、硬件，但是用户使用云计算服务的质量依赖云计算本身的质量
可维护性	提供专业的软件管理和维护服务，减少了普通用户软件平台的日常维护管理成本	是否所有的软件应用都适合在云计算环境下开发应用，而以往的软件应用如何移植到云计算环境下
交互性	用户可以根据业务需要动态地按需请求云计算服务，处理高峰期负载并在非高峰期释放资源	云计算服务提供商的实际扩展能力有限，需要多个云计算服务商间的交互，而云计算服务之间的交互性较差

3. 云计算应用分类

云计算的类型从不同的角度有不同的划分，本节在横向上按部署方式，在纵向上按云计算从底层到高层提供服务的方式分类介绍各种云计算，结合典型的云计算服务平台，由此在图 3-3 中分析云计算框架的构成，讨论各层次需要构建的机制和实现方案。

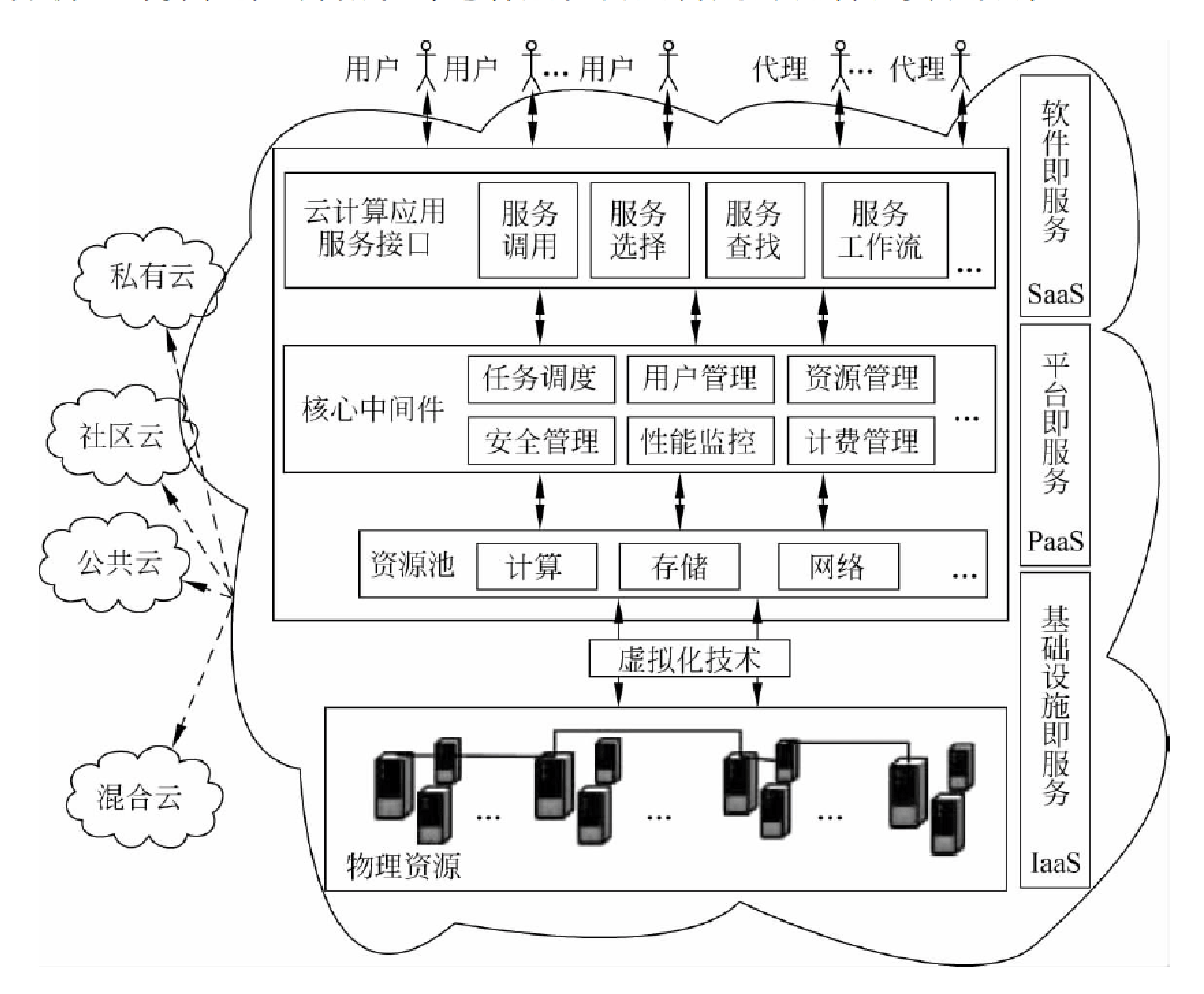


图 3-3　云计算框架图

从云计算部署的角度，云计算分为私有云、社区云、公共云和混合云。私有云被一个组织管理操作。社区云由多个组织共同管理操作，具有一致的任务调度和安全策略。公共云由一个组织管理维护，提供对外的云服务，可以被公众所拥有。混合云是以上两种

或两种以上云的组合。从云计算服务的角度，云计算服务类型可以分为基础设施即服务(Infrastructure as a Service，IaaS)、平台即服务(Platform as a Service，PaaS)、软件即服务(Software as a Service，SaaS)。

(1) IaaS 在服务层次上是底层服务，接近物理硬件资源，通过虚拟化的相关技术，为用户提供计算、存储、网络以及其他资源方面的服务，以便用户能够部署操作系统和运行软件。这一层典型的服务如亚马逊的弹性云(Amazon，EC2)。EC2 与 Google 提供的云计算服务不同，Google 只为互联网上的应用提供云计算平台，开发人员无法在这个平台上工作，因此只能转而通过开源的 Hadoop 软件支持来开发云计算应用。而 EC2 给用户提供一个虚拟的环境，使得可以基于虚拟的操作系统环境运行自身的应用程序。同时，用户可以创建亚马逊机器镜像(AMI)，镜像包括库文件、数据和环境配置，通过弹性计算云的网络界面去操作在云计算平台上运行的各个实例(Instance)，同时用户需要为相应的简单存储服务(S3)和网络流量付费。

(2) PaaS 是构建在基础设施即服务之上的服务，用户通过云服务提供的软件工具和开发语言，部署自己需要的软件运行环境和配置。用户不必控制底层的网络、存储、操作系统等技术问题，底层服务对用户是透明的，这一层服务是软件的开发和运行环境。这一层服务是一个开发、托管网络应用程序的平台，代表性的有 Google App Engine 和 Microsoft Azure。使用 Google App Engine，用户将不再需要维护服务器，用户基于 Google 的基础设施上传、运行应用程序软件。目前，Google App Engine 用户使用一定的资源是免费的，如果使用更多的带宽、存储空间等需要另外收取费用。Google App Engine 提供一套 API 使用 Python 或 Java 来方便用户编写可扩展的应用程序，但仅限 Google App Engine 范围的有限程序，现存很多应用程序还不能很方便地运行在 Google App Engine 上。Microsoft Azure 构建在 Microsoft 数据中心内，允许用户应用程序，同时提供了一套内置的有限 API，方便开发和部署应用程序。此平台包含在线服务 Live Service、关系数据库服务 SQL Services、各式应用程序服务器服务 NET Services 等。

(3) SaaS 是前两层服务所开发的软件应用，不同用户以简单客户端的方式调用该层服务，例如以浏览器的方式调用服务。用户可以根据自己的实际需求，通过网络向提供商定制所需的应用软件服务，按服务多少和时间长短支付费用。最早提供该服务模式的是 Saleforce 公司运行的客户关系管理(CRM)系统，它是在该公司 PaaS 层 force. com 平台之上开发的 SaaS。Google 的在线办公软件如文档、表格、幻灯片处理也采用 Saas 服务模式。

云计算提供的不同层次服务使开发者、服务提供商、系统管理员和用户面临许多挑战。图 3-3 对此做出了归纳概述。底层的物理资源经过虚拟化转变为多个虚拟机，以资源池多重租赁的方式提供服务，提高了资源的效用。核心中间件起到任务调度、资源和安全管理、性能监控、计费管理等作用。一方面，云计算服务涉及大量的调用第三方软件及框架和重要数据处理的操作，这需要有一套完善的机制，以保证云计算服务安全有效地运行；另一方面，虚拟化的资源池所在的数据中心往往电力资源耗费巨大，解决这样的问题需要设计有效的资源调度策略和算法。在用户通过代理或者直接调用云计算服务的时候，需要和服务提供商之间建立服务等级协议(Service Level Agreement，SLA)，那么必然需要服务性能监控，以便设计出比较灵活的付费方式。此外，还需要设计便捷的应用接口，方便服务调用。而用户在调用中选择什么样的云计算服务，这就要设计合理的度量标准并建立一个全球云

计算服务市场以供选择调用。

3.1.3 数据中心

前面说到计算机的发展经历了几个阶段，从早期的超级计算机到 PC 时代，再到互联网 dot-com 时代，然后进入了现今的云计算时代。最早期的计算机系统操作和维护都复杂，需要一个特殊的环境来操作。同时安全非常重要，因为计算机非常贵，并且常常被用于军事目的。因此除了机房的设计和装修，控制机房的访问权限也都考虑在列。随着微型计算机的普及，在 20 世纪 90 年代一些微型计算机(被称为服务器)逐步在一些公司的机房得到使用，并且机房的规模逐步扩大。

到了 dot-com 时代，数据中心在全球取得了快速的发展。很多的互联网公司需要不停地增加服务器，并具备快速的 Internet 连接。一些公司开始建立大型的计算机服务机房，被称为 Internet 数据中心(IDCs)，它提供了商业的系统部署和操作的解决方案，为这些互联网公司提供专业化的基础设施服务。自 2007 年起，数据中心的设计、构建和运营逐步形成了了一门学科，并且有国际化的标准组织，如电讯产业联合会(ITU)，详细制定数据中心相关的需求和标准。

维基百科给出的数据中心定义是"一整套复杂的设施，它不仅包括计算机、系统和其他与之配套的设备(例如通信和存储系统)，还包含冗余的数据通信连接、环境控制设备、监控设备以及各种安全装置"。目前，数据中心在各行业都发挥着至关重要的作用，承载着企业的关键业务，为用户提供及时可靠的数据存储、数据检索、数据分析及发掘、高性能计算等服务，如 Google 数据中心为全球网民提供搜索、视频等服务，腾讯的数据中心为用户提供微信、QQ、游戏等服务。

从数据中心模式服务的发展而言，其产生和演化经历了三个阶段：主机共享时期，主机托管时期，应用服务托管时期。起初就是主机托放服务，只为用户提供电源、带宽，机器重新启动都要自己来做；随后出现了主机托管服务，主要是带宽上有保证，电源上有备份，并且可以部分代为管理；一些大型的客户要求更多的增值服务，包括一些关键性业务，如要求安全性、数据流的分析、资源的占用状况等，需求越来越多，要求有更多的服务。在这种情况下，出现了提供综合服务的大型数据中心服务商。这个时期比较成熟的数据中心模式才算正式出现。

随着云计算的发展，IT 资源的应用和共享方式发生了巨大的变化。云计算是网格计算、并行计算、分布式计算、虚拟化、负载均衡等传统计算机和网络技术发展融合的产物。它是一种全新的计算方式和资源使用方式，普通用户可以十分方便地接入强大的 IT 资源并按需部署自己的服务，同时多种全新的业务模式能够得以实现，另外 IT 资源和服务能够从底层基础设施中抽象出来，这极大增强了资源的共享性和灵活性。数据中心是云计算的实现平台，云计算时代的数据中心已经从原本的数据存储节点转变为面向服务和应用的 IT 核心节点。随着各种数据密集型业务的出现，数据中心已经成为唯一能够支持大规模云计算应用的服务平台(例如 Microsoft Azure、Amazon EC2、Google Search、Facebook 等)。同时，为了给云计算提供"无限可能"的资源池，数据中心必须包含更多存储资源、计算资源以及通信带宽。新一代数据中心将包含数万乃至数十万台服务器，例如，目前 Google 在全球有三十多个大型数据中心，单个数据中心服务器数目超过了 45 000 台，微软在印第安纳州

建立的数据中心投资规模达 6.7 亿美元，在计划构建的数据中心可容纳的服务器数目高达 300 000 台，国内的大型互联网公司如阿里巴巴、腾讯新建的数据中心规模也都超过 200 000 台服务器。

3.2　数据采集

足够的数据量是企业大数据战略建设的基础，因此数据采集是大数据价值挖掘中的重要的一环，其后的分析挖掘都建立在数据采集的基础上。

数据的采集有基于物联网传感器的采集，也有基于网络信息的数据采集。比如在智能交通中，数据的采集有基于 GPS 的定位信息采集、基于交通摄像头的视频采集、基于交通卡口的图像采集、基于路口的线圈信号采集等。而在互联网上的数据采集是对各类网络媒介，如搜索引擎、新闻网站、论坛、微博、博客、电商网站等的各种页面信息和用户访问信息进行采集，采集的内容主要有文本信息、URL、访问日志、日期和图片等。之后需要把采集到的各类数据进行清洗、过滤、去重等各项预处理并分类归纳存储。

在分布式系统中，经常需要采集各个节点的日志，然后进行分析。在数据量呈爆炸式增长的今天，数据的种类丰富多样，也有越来越多的数据需要将存储和计算放到分布式平台。数据采集过程中的 ETL 工具将分布的、异构数据源中的不同种类和结构的数据抽取到临时中间层后进行清洗、转换、分类、集成，最后加载到对应的数据存储系统，如数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。企业每天都会产生大量的日志数据，对这些日志数据的处理需要特定的日志系统。因为与传统的数据相比，大数据的体量巨大，产生速度非常快，对数据的预处理需要实时快速，因此在 ETL 的架构和工具选择上，也需要采用分布式内存数据、实时流处理系统等现代信息技术。

3.2.1　系统日志采集方法

很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如 Hadoop 的 Chukwa，Cloudera 的 Flume，LinkedIn 的 Kafka，Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足每秒数百 MB 的日志数据采集和传输需求。

3.2.2　网络数据采集方法：对非结构化数据的采集

网络数据采集是指通过网络爬虫或网站公开 API 等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

除了网络中包含的内容之外，对于网络流量的采集可以使用 DPI 或 DFI 等带宽管理技术进行处理。

3.2.3　其他数据采集方法

对于企业生产经营数据或学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

下面对系统日志采集的工具进行详细介绍。

1. Facebook Scribe

Scribe 是 Facebook 开源的日志收集系统，在 Facebook 内部已经得到大量的应用。它能够从各种日志源上收集日志，存储到一个中央存储系统(可以是 NFS、分布式文件系统等)上，以便于进行集中统计分析处理。它为日志的“分布式收集，统一处理”提供了一个可扩展的、高容错的方案。

Scribe 最重要的特点是容错性好。当后端的存储系统 crash 时，Scribe 会将数据写到本地磁盘上，当存储系统恢复正常后，Scribe 再将日志重新加载到存储系统中。

Scribe 架构如图 3-4 所示。

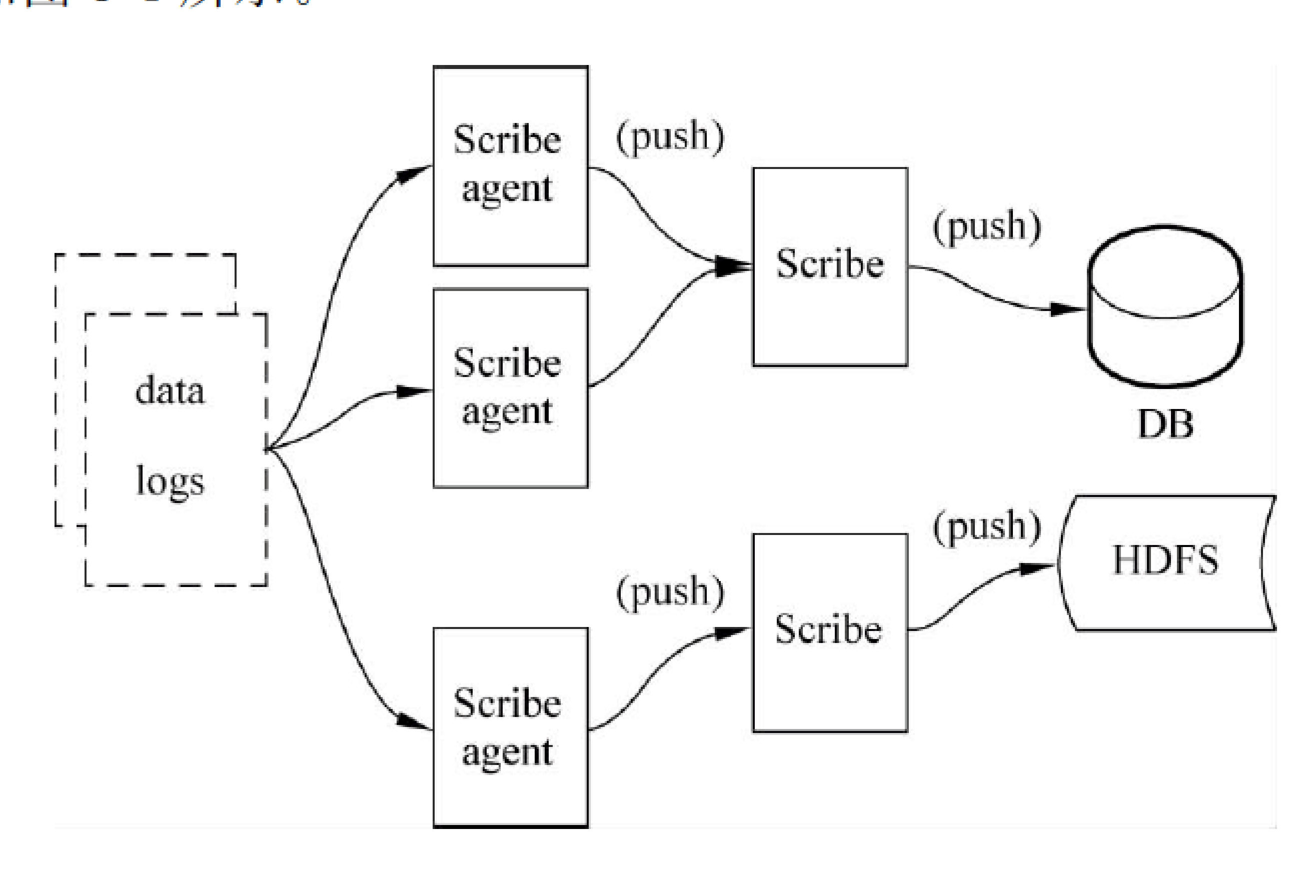


图 3-4　Scribe 采集架构

Scribe 的架构比较简单，主要包括三部分，分别为 Scribe agent、Scribe 和存储系统。

(1) Scribe agent。Scribe agent 实际上是一个 thrift client。向 Scribe 发送数据的唯一方法是使用 thrift client，Scribe 内部定义了一个 thrift 接口，用户使用该接口将数据发送给 Server。

(2) Scribe。Scribe 接收到 thrift client 发送过来的数据，根据配置文件，将不同 topic 的数据发送给不同的对象。Scribe 提供了各种各样的 store，如 file、HDFS 等，Scribe 可将数据加载到这些 store 中。

(3) 存储系统。存储系统实际上就是 Scribe 中的 store，当前 Scribe 支持非常多的 store，包括 file(文件)，buffer(双层存储，一个主储存，一个副存储)，network(另一个 Scribe 服务器)，bucket(包含多个 store，通过 hash 将数据存到不同 store 中)，null(忽略数据)，thriftfile(写到一个 Thrift FileTransport 文件中)和 multi(把数据同时存放到不同 store 中)。

2. Apache Chukwa

Chukwa 是一个非常新的开源项目，由于其属于 Hadoop 系列产品，因而使用了很多 Hadoop 的组件(用 HDFS 存储，用 MapReduce 处理数据)，它提供了很多模块以支持 Hadoop 集群日志分析，如图 3-5 所示。

1) 需求

(1) 灵活的、动态可控的数据源；

(2) 高性能、高可扩展的存储系统；

(3) 合适的框架，用于对收集到的大规模数据进行分析。

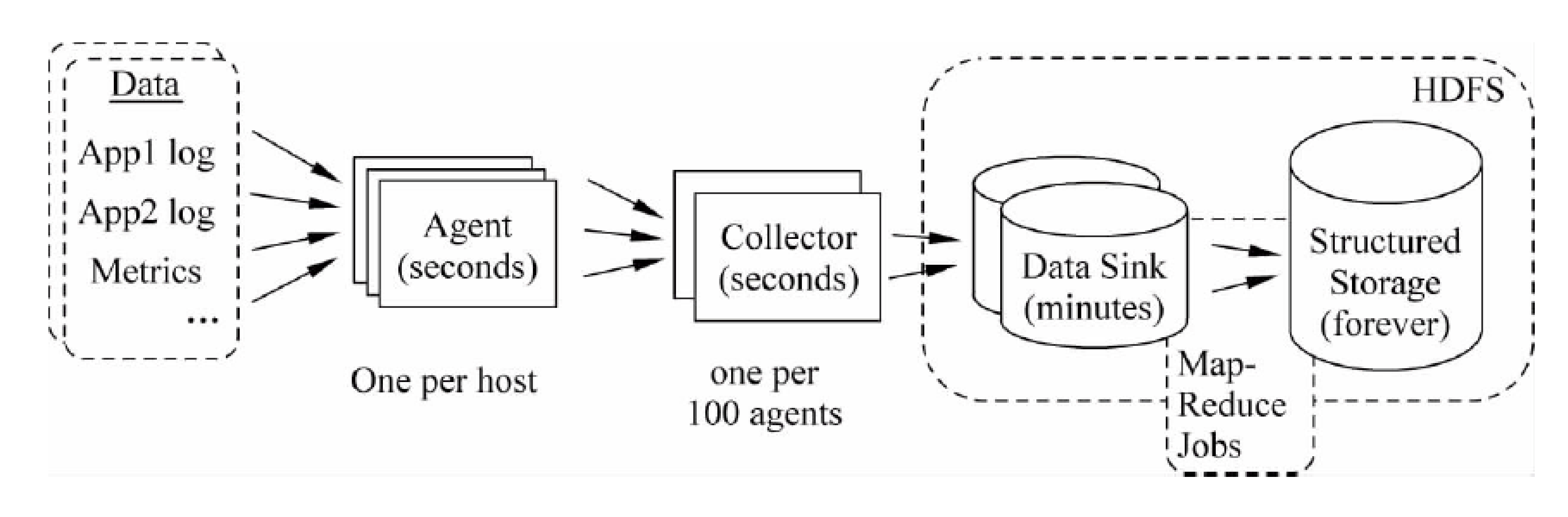


图 3-5　Chukwa 采集架构

2）架构

Chukwa 中主要有三种角色，分别为：Adaptor，Agent，Collector。

（1）Adaptor 数据源。可封装其他数据源，如 file、UNIX 命令行工具等。

目前可用的数据源有：Hadoop logs，应用程序度量数据，系统参数数据（如 Linux CPU 使用流率）。

（2）HDFS 存储系统。Chukwa 采用了 HDFS 作为存储系统。HDFS 的设计初衷是支持大文件存储和小并发高速写的应用场景，而日志系统的特点恰好相反，它需支持高并发低速率的写和大量小文件的存储。需要注意的是，直接写到 HDFS 上的小文件是不可见的，直到关闭文件。另外，HDFS 不支持文件重新打开。

（3）Collector 和 Agent。为了克服（2）中的问题，增加了 Agent 和 Collector 阶段。

Agent 的作用：给 Adaptor 提供各种服务，包括启动和关闭 Adaptor，将数据通过 HTTP 传递给 Collector；定期记录 Adaptor 状态，以便 crash 后恢复。

Collector 的作用：对多个数据源发过来的数据进行合并，然后加载到 HDFS 中；隐藏 HDFS 实现的细节，如 HDFS 版本更换后，只需修改 Collector 即可。

（4）Demux 和 Achieving。直接支持利用 MapReduce 处理数据。它内置了两个 MapReduce 作业，分别用于获取 data 和将 data 转化为结构化的 log。存储到 data store（可以是数据库或者 HDFS 等）中。

3. LinkedIn Kafka

Kafka 是 2010 年 12 月开源的项目，采用 Scala 语言编写，使用了多种效率优化机制，整体架构（如图 3-6 所示）比较新颖（Push/Pull），更适合异构集群。

1）设计目标

（1）数据在磁盘上的存取代价为 $O(1)$。

（2）高吞吐率，在普通的服务器上每秒也能处理几十万条消息。

（3）分布式架构，能够对消息分区。

（4）支持将数据并行地加载到 Hadoop。

2）架构

Kafka 实际上是一个消息发布订阅系统。Producer 向某个 topic 发布消息，而 Consumer 订阅某个 topic 的消息，进而一旦有新的关于某个 topic 的消息，Broker 会传递给订阅它的所有 Consumer。在 Kafka 中，消息是按 topic 组织的，而每个 topic 又会分为多个 partition，这样便于管理数据和进行负载均衡。同时，它也使用了 Zookeeper 进行负载均衡。

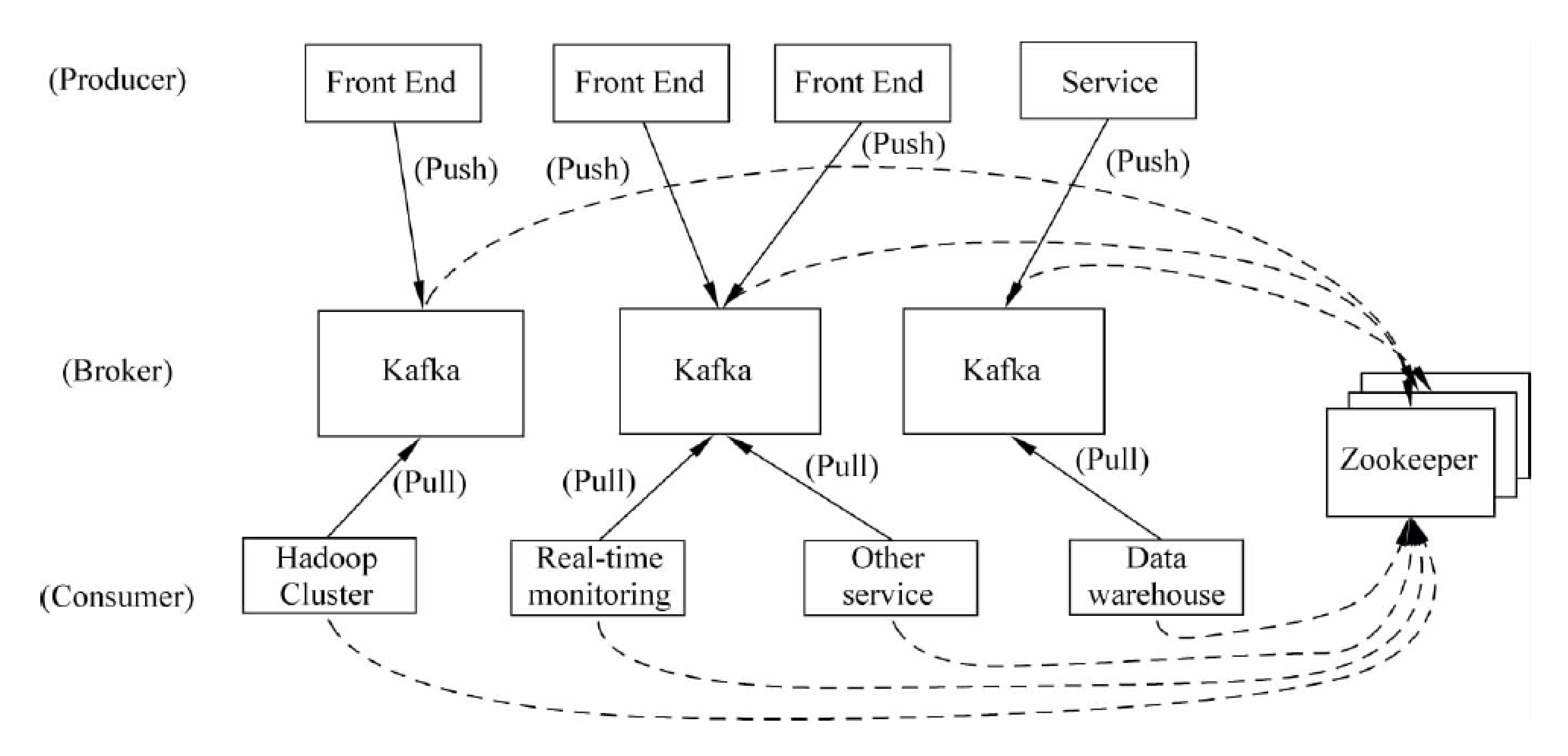


图 3-6 Kafka 采集处理架构

Kafka 中主要有三种角色，分别为 Producer、Broker 和 Consumer。

(1) Producer。Producer 的任务是向 Broker 发送数据。Kafka 提供了两种 Producer 接口，一种是 low level 接口，使用该接口会向特定的 Broker 的某个 topic 下的某个 partition 发送数据；另一种是 high level 接口，该接口支持同步/异步发送数据，基于 Zookeeper 的 Broker 自动识别和负载均衡(基于 Partitioner)。

其中，基于 Zookeeper 的 Broker 自动识别值得一说。Producer 可以通过 Zookeeper 获取可用的 Broker 列表，也可以在 Zookeeper 中注册 Listener，该 Listener 在以下情况下会被唤醒：①添加一个 Broker；②删除一个 Broker；③注册新的 topic；④Broker 注册已存在的 topic。

当 Producer 得知以上事件时，可根据需要采取一定的行动。

(2) Broker。Broker 采取了多种策略提高数据处理效率，包括 sendfile 和 zero copy 等技术。

(3) Consumer。Consumer 的作用是将日志信息加载到中央存储系统上。Kafka 提供了两种 Consumer 接口，一种是 low level 的，它维护到某一个 Broker 的连接，并且这个连接是无状态的，即每次从 Broker 上 Pull 数据时，都要告诉 Broker 数据的偏移量。另一种是 high level 接口，它隐藏了 Broker 的细节，允许 Consumer 从 Broker 上 Push 数据而不必关心网络拓扑结构。更重要的是，对于大部分日志系统而言，Consumer 已经获取的数据信息都由 Broker 保存，而在 Kafka 中，由 Consumer 自己维护所取数据信息。

4. Cloudera Flume

Flume 是 Cloudera 于 2009 年 7 月开源的日志系统。它内置的各种组件非常齐全，用户几乎不必进行任何额外开发即可使用，如图 3-7 所示是 Flume 采集架构。

1) 设计目标

(1) 可靠性。当节点出现故障时，日志能够被传送到其他节点上而不会丢失。Flume 提供了三种级别的可靠性保障，从强到弱依次分别为：end-to-end(收到数据 agent 首先将 event 写到磁盘上，当数据传送成功后，再删除；如果数据发送失败，可以重新发送)，Store

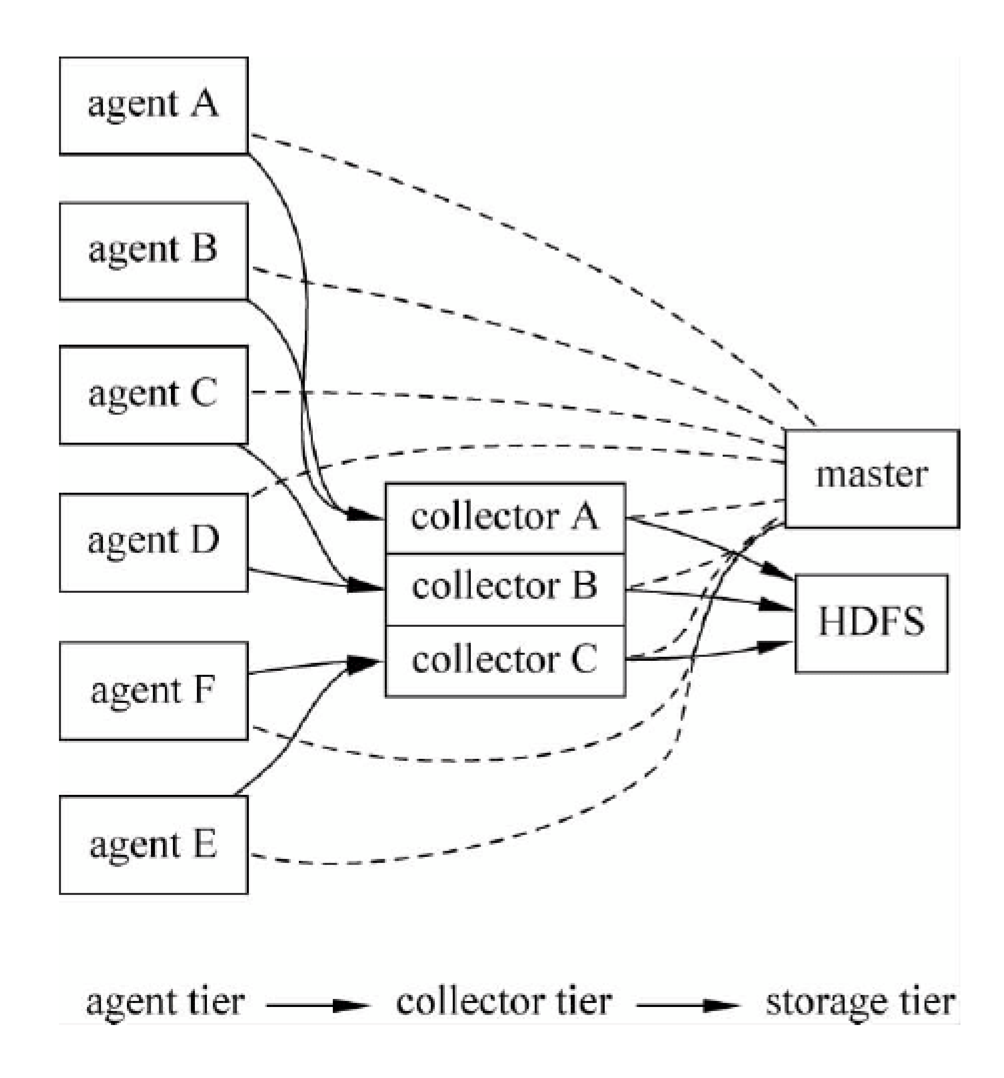


图 3-7　Flume 采集架构

on failure(这也是 Scribe 采用的策略，当数据接收方 crash 时，将数据写到本地，待恢复后继续发送)，Best effort(数据发送到接收方后，不会进行确认)。

(2) 可扩展性。Flume 采用了三层架构，分别为 agent，collector 和 storage，每一层均可以水平扩展。其中，所有 agent 和 collector 由 master 统一管理，这使得系统容易监控和维护，且 master 允许有多个(使用 Zookeeper 进行管理和负载均衡)，这就避免了单点故障问题。

(3) 可管理性。所有 agent 和 colletor 由 master 统一管理，这使得系统便于维护。用户可以在 master 上查看各个数据源或者数据流执行情况，且可以对各个数据源配置和动态加载。Flume 提供了 Web 和 shell script command 两种形式对数据流进行管理。

(4) 功能可扩展性。用户可以根据需要添加自己的 agent、colletor 或者 storage。此外，Flume 自带了很多组件，包括各种 agent(file，syslog 等)、collector 和 storage(file，HDFS 等)。

2) 架构

正如前面提到的，Flume 采用了分层架构，由三层组成，分别为 agent、collector 和 storage。其中，agent 和 collector 均由两部分组成：source 和 sink，source 是数据来源，sink 是数据去向。

(1) agent。agent 的作用是将数据源的数据发送给 collector，Flume 自带了很多直接可用的数据源(source)。

(2) collector。collector 的作用是将多个 agent 的数据汇总后，加载到 storage 中。它的 source 和 sink 与 agent 类似。

(3) storage。storage 是存储系统，可以是一个普通文件，也可以是 HDFS，Hive，HBase 等。

3.3　数据存储

云计算中的数据存储是实现云计算系统架构中的一个重要组成部分。云存储专注于解决云计算中海量数据的存储问题，它既可以给云计算技术提供专业的存储解决方案，又可以

独立发布存储服务。云存储将存储作为服务，它将分别位于网络中不同位置的大量类型各异的存储设备通过集群应用、网格技术和分布式文件系统等集合起来协同工作，通过应用软件进行业务管理，并通过统一的应用接口对外提供数据存储和业务访问功能。目前，云存储的兴起正在颠覆传统的存储系统架构，其正以良好的可扩展性、性价比和容错性等优势得到业界的广泛认同。云存储系统具有良好的可扩展性、容错性，以及内部实现对用户透明等特性，这一切都离不开分布式文件系统的支撑。现有的云存储分布式文件系统包括 Google GFS、Hadoop HDFS、Lustre、FastDFS、Clemsom 大学的 PVFS、Sun PFS、加州大学 Santa Cruz 分校 Sage Weil 设计的 Ceph 和 Taobao TFS 等。

目前存在的数据库存储方案有结构化存储方案 SQL、非结构化存储方案 NoSQL 和革新的结构化方案 NewSQL。

SQL 一般指的是关系型数据库 RDBMS，是目前为止企业应用中最为成功的数据存储方案，仍有相当大一部分的企业把 SQL 数据库作为数据存储方案。关系型数据库能够较好地保证事务的 ACID 特性，但在可扩展性、可用性等方面，表现出较大的不足，并且只能处理结构化的数据，面对数据的多样性、处理数据的实时性等方面，都不能满足大数据时代环境下数据处理的需要。使用较多的 SQL 产品有 IBM DB2、Oracle、MySQL、MS SQL Server 等。

NoSQL 是为了解决 SQL 的不足而产生的。大数据时代，数据的显著特点就是数据量大，这些数据是 TB 或 PB 级别以上的量级；数据结构不统一，包括结构化的、半结构化的和非结构化的数据，其规模或复杂程度超出了常用传统数据库和软件技术所能管理和处理的数据集范围。

NoSQL 有良好、便捷的横向扩展性，可以满足海量数据的存储需求。NoSQL 是一种无模式的数据存储模型，可以应对 Web 应用上各种半结构化的数据，灵活简单的数据模型以及弱一致性的特性使得高并发情况下数据查询的性能优异。可以说 NoSQL 是大数据时代数据库领域不可或缺的重要一员。NoSQL 的主要优势与特点如下。

(1) 灵活的数据模型：多样的数据模型支持，有基于 key-value 的、基于列存储的、基于图的一系列数据模型。

(2) 灵活的可扩展性、经济性：相对于 RDBMS 来说，NoSQL 最突出的一个特点就是横向扩展，NoSQL 数据库通常使用廉价的服务器集群来管理膨胀的数据和事务数量，而 RDBMS 通常需要依靠昂贵的专有服务器和存储系统来做到这一点。使用 NoSQL，每 GB 的成本或每秒处理事务的成本，都比使用 RDBMS 少很多倍，可以花费更低的成本来存储和处理更多的数据。

NoSQL 能取得高扩展性是因为在设计时放松了事务的 ACID 特性。根据 CAP 定理，数据库系统不可能同时满足一致性(Consistency)、可用性(Availability)和分区容错性(Partition Tolerance)三个特性，最多只能选择其中的两项。NoSQL 数据库在设计时经常会保证分区容错性，而牺牲一致性或可用性，因而 NoSQL 的应用范围也受到了很大的限制。如何构建具有高可扩展性、高可用性、高性能的，同时还能保证 ACID 事务特性的数据库就成为新的发展方向。现有的 NoSQL 数据库有很多，例如 HBase、Cassandra、MongoDB、CouchDB、Hypertable、Redis 等。

NewSQL 是为解决上述数据库存在的不足，顺应科技发展的产物。该类数据库要求，

不仅要具有 NoSQL 对海量数据的存储管理能力，还要保持对传统数据库支持 ACID 和 SQL 等特性。目前，NewSQL 系统产品有 H-Store、VoltDB、NuoDB、TokuDB、MemSQL 等。

3.3.1 结构化数据存储

结构化数据即行数据，存储在数据库里，可以用二维表结构来逻辑表达现实的数据。传统的关系型数据库存储的都是结构化数据。常用的关系型数据库有：Oracle Database、MySQL、SQL Server 和 DB2 等。下面简单介绍下这些数据库。

1. Oracle Database

Oracle 数据库系统是美国 Oracle 公司（甲骨文）提供的以分布式数据库为核心的一组软件产品，是目前最流行的客户/服务器（Client/Server）或 B/S 体系结构的数据库之一。比如 SilverStream 就是基于数据库的一种中间件。Oracle 数据库是目前世界上使用最为广泛的数据库管理系统，作为一个通用的数据库系统，它具有完整的数据管理功能；作为一个关系数据库，它是一个完备关系的产品；作为分布式数据库，它实现了分布式处理功能。但只要在一种机型上学习了 Oracle 知识，便能在各种类型的机器上使用它。

Oracle 数据库最新版本为 Oracle Database 12c。Oracle 数据库 12c 引入了一个新的多承租方架构，使用该架构可轻松部署和管理数据库云。此外，一些创新特性可最大限度地提高资源使用率和灵活性，如 Oracle Multitenant 可快速整合多个数据库，而 Automatic Data Optimization 和 Heat Map 能以更高的密度压缩数据和对数据分层。这些独一无二的技术进步再加上在可用性、安全性和大数据支持方面的主要增强，使得 Oracle 数据库 12c 成为私有云和公有云部署的理想平台。

2. MySQL

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统，在 Web 应用方面 MySQL 是最好的 RDBMS（Relational Database Management System，关系数据库管理系统）应用软件之一。

MySQL 是一种关联数据库管理系统，关联数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。

MySQL 所使用的 SQL 是用于访问数据库的最常用标准化语言。MySQL 软件采用了双授权政策，分为社区版和商业版，由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点，一般中小型网站的开发都选择 MySQL 作为网站数据库。

由于其社区版的性能卓越，搭配 PHP 和 Apache 可组成良好的开发环境。

3. SQL Server

SQL Server 是一个关系数据库管理系统。它最初是由 Microsoft、Sybase 和 Ashton-Tate 三家公司共同开发的，于 1988 年推出了第一个 OS/2 版本。在 Windows NT 推出后，Microsoft 与 Sybase 在 SQL Server 的开发上就分道扬镳了，Microsoft 将 SQL Server 移植到 Windows NT 系统上，专注于开发推广 SQL Server 的 Windows NT 版本。Sybase 则较专注于 SQL Server 在 UNIX 操作系统上的应用。

4. DB2

IBM DB2是美国IBM公司开发的一套关系型数据库管理系统，它主要的运行环境为UNIX（包括IBM自家的AIX）、Linux、IBM i（旧称OS/400）、z/OS，以及Windows服务器版本。

DB2主要应用于大型应用系统，具有较好的可伸缩性，可支持从大型计算机到单用户环境，应用于所有常见的服务器操作系统平台下。DB2提供了高层次的数据利用性、完整性、安全性、可恢复性，以及小规模到大规模应用程序的执行能力，具有与平台无关的基本功能和SQL命令。DB2采用了数据分级技术，能够使大型计算机数据很方便地下载到LAN数据库服务器，使得客户/服务器用户和基于LAN的应用程序可以访问大型计算机数据，并使数据库本地化及远程连接透明化。DB2以拥有一个非常完备的查询优化器而著称，其外部连接改善了查询性能，并支持多任务并行查询。DB2具有很好的网络支持能力，每个子系统可以连接十几万个分布式用户，可同时激活上千个活动线程，对大型分布式应用系统尤为适用。

DB2除了可以提供主流的OS/390和VM操作系统，以及中等规模的AS/400系统之外，IBM还提供了跨平台（包括基于UNIX的Linux，HP-UX，Sun Solaris等；还有用于个人计算机的OS/2操作系统，以及微软的Windows和其早期的系统）的DB2产品。DB2数据库可以通过使用微软的开放数据库连接（ODBC）接口、Java数据库连接（JDBC）接口，或者CORBA接口代理被任何的应用程序访问。

3.3.2 非结构化数据存储

相对于结构化数据（即行数据，存储在数据库里，可以用二维表结构来逻辑表达现实的数据）而言，不方便用数据库二维逻辑表来表现的数据即称为非结构化数据，包括所有格式的办公文档、文本、图片、标准通用标记语言下的子集XML、HTML、各类报表、图像和音频/视频信息等。

非结构化数据库是指其字段长度不等，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库，用它不仅可以处理结构化数据（如数字、符号等信息）而且更适合处理非结构化数据（全文文本、图像、声音、影视、超媒体等信息）。

非结构化Web数据库主要是针对非结构化数据而产生的，与以往流行的关系数据库相比，其最大区别在于它突破了关系数据库结构定义不易改变和数据定长的限制，支持重复字段、子字段以及变长字段并实现了对变长数据和重复字段进行处理和数据项的变长存储管理，在处理连续信息（包括全文信息）和非结构化信息（包括各种多媒体信息）中有着传统关系型数据库所无法比拟的优势。

为了解决非结构化数据的存储和并发计算以及扩展能力，NoSQL登上了舞台，如Google的BigTable，Amazon的Dynamo，以及Apache的HBase。NoSQL支持强大的水平扩展能力和高性能，与关系数据库不同的是，NoSQL可以采用松弛一致性。如最终一致性，或交易仅限于单个的数据项。像在Dynamo中为了提供高的写的能力（购物时不会因为并发而导致添加购物车不成功，影响用户体验），不得不采取最终一致性。在Dynamo中牺牲了一致性，但是提供高的可用性。另外，Dynamo采用非集中化管理，使得每个节点都是同等地位，充分利用分布式哈希表（DHT）的一种实现即一致性哈希，使得Dynamo提供强大

的可扩展性。HBase 可以说提供强的一致性，但是牺牲掉了一定的高可用性，比如存在单点故障，在当一个 Region Server 出问题或失去联系时，需要 master 来重新部署原 Region Server 下面的 Region 到别的空闲服务器下。这段时间无法与下面的 Region 联系。HBase 是 Apache 的顶级项目 Hadoop 的一个组成部分，Hadoop 是一种分布式系统基础架构。它可以充分利用集群的威力高速运算和存储。后文会着重介绍 HBase。

3.4 数据处理

在大数据的环境下，数据处理除了标准的查询、统计、分析、商业智能之外，主要还体现在数据挖掘、深度学习、社交计算、计算广告等几个方面。

数据挖掘又称从数据库中发现知识(KDD)、数据分析、数据融合以及决策支持。数据挖掘领域已经有了较长时间的发展，但随着研究的不断深入、应用的愈发广泛，数据挖掘的关注焦点也逐渐有了新的变化。其总的趋势是数据挖掘研究和应用更加“大数据化”和“社会化”。在用户层面，移动计算设备的普及与大数据革命带来的机遇使得搜索引擎对用户所处的上下文环境具有了前所未有的深刻认识，但对于如何将认识上的深入转化为用户信息获取过程的便利仍然缺乏成功经验。近年来，以用户个性化、用户交互等为代表的研究论文的数量大幅增加。除此之外，社交网络服务的兴起对互联网数据环境和用户群体均将形成关键性的影响，如何更好地面对相对封闭的社交网络数据环境和被社交关系组织起来的用户群体，也是数据挖掘面临的机遇与挑战。

深度学习是机器学习研究中的一个新的领域。它在于建立模拟人脑进行分析学习的神经网络，模仿人脑机制来解释一些特定类别的数据，例如图像、语音和文本。它是无监督学习的一种。深度学习的主要思想是增加神经网络中隐藏层的数量，使用大量的隐藏层来增强神经网络对特征筛选的能力，以增加网络层数的方式来取代之前依赖人工技巧的参数调优，从而能够用较少的参数表达出复杂的模型函数，从而逼近机器学习的终极目标——知识的自动发现。

社交网络每天都会产生大量的用户数据，它吸引着无数研究者从无序的数据中发掘有价值的信息。在社交网络的分析与研究过程中，会利用到社会学、心理学甚至是医学的基本理论来作为指导。社交网络上的传播模型、虚假信息和机器人账号的识别，基于社交网络信息对股市、大选以及传染病的预测，社区圈子的区别，社交网络中人物的影响力等，都可以作为社交网络中的研究课题。通过人工智能领域的机器学习、图论等算法对社交网络中行为和未来的趋势进行模拟和预测。

计算广告是一门正在兴起的分支学科。它由信息科学、统计学、计算机科学以及微观经济学等学科交叉融合而成。它涉及大规模搜索和文本分析、信息获取、统计模型、机器学习、分类、优化及微观经济学。计算广告学所面临的最主要挑战是在特定语境下特定用户和相应的广告之间找到“最佳匹配”。语境可以是用户在搜索引擎中输入的查询词，也可以是用户正在读的网页，还可以是用户正在看的电影等。而用户相关的信息可能非常多也可能非常少。潜在广告的数量可能达到几十亿。因此，取决于对“最佳匹配”的定义，面临的挑战可能导致在复杂约束条件下的大规模优化和搜索问题。

面向大数据处理的数据查询、统计、分析、挖掘等需求，促生了大数据的不同计算模式。我们将大数据的计算模式按照时间维度和数据处理方式两个方式来进行划分。从时间维度

上来讲,可以分为实时计算和离线(非实时)计算。

实时计算,强调的是计算能够实时完成。这里的实时,并没有严格的定义,一般都跟应用的需求有关。在大数据处理领域,一般指的是处理时间在秒级,在一些对响应时间要求很严格的工业级应用中,要求甚至达到毫秒级。

离线计算,则与实时计算相反,对处理时间没有强制要求,但一般计算数据量会相当大,处理时间能达到几个小时甚至几天。

从数据的处理方式来说,大数据处理可以分为流计算和批处理。

流计算,指的是数据在源源不断产生,并且数据一到来就立即进行处理的计算模式,该模式一般会一直占用计算资源不进行释放,从而保证数据到来时能够马上进行处理。流计算具有如下的特点。

(1) 类似数学中的连续函数,计算在连续进行;

(2) 并不保证计算是实时的,它只保证数据在第一时间被处理;

(3) 资源的持续占用。

批处理,指数据到来后,并不是立即处理,而是累积到一定量才进行处理。因此,该模式不要求对计算资源的持续占用。相对于流处理,批处理的特点如下。

(1) 类似数学中的离散函数,计算在每个离散点进行;

(2) 批处理并不意味着计算一定达不到实时,它只说明数据是以批量的形式处理;

(3) 不用一直占用资源。

传统技术通过缩短批处理间隔时间可以实现准实时计算。传统技术中,大多采用批处理模式对数据进行处理。为了达到实时效果,采用不断缩短批处理间隔时间的方式来实现实时计算。例如实时数据库技术,缩小批处理数据累积时间,从小时转为分钟等,并提高机器处理性能,就能实现准实时计算。

但是,随着数据量的增多,且业界对实时间隔时间的定义越来越短,批处理数据累积时间也越来越短,甚至直接使累积时间为0,这样,流计算的原型就诞生了。

流计算一般是为实时计算场景所设计。由于流计算的产生本来就是源于实时计算的需求,因此现有的流计算技术均采用了内存计算、并行计算等多种计算技术,提高了快速实时计算能力,所以流计算能够解决实时计算问题。

但从流计算本质来看,如果一个系统能保证数据进入系统时就开始处理,但是,整个处理过程可能由于某些高延迟性操作如大量磁盘读写操作,导致处理时间较长,该系统依然是流计算,而不是实时计算。

在实际的大数据处理场景中,一般不存在流计算和离线计算相结合的场景,因此在整体上,我们把大数据的计算模式只分为离线批处理、实时交互计算和流计算三种模式。

3.4.1 离线批处理

随着云计算技术的广泛应用和发展,基于开源的 Hadoop 分布式存储系统和 MapReduce 数据处理模式的分析系统也得到了广泛的应用。Hadoop 采用数据分块及自恢复机制,能支持 PB 级的分布式的数据存储,而且它是基于 MapReduce 分布式处理模式对这些数据进行分析和处理的。MapReduce 编程模型可以很容易地将多个通用批数据处理任务和操作在大规模集群上并行化,而且它有自动化的故障转移功能。MapReduce 编程模

型在 Hadoop 这样的开源软件的带动下被广泛采用，如在 Web 搜索、欺诈检测等各种各样的实际应用中。

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，而且是以一种可靠、高效、可伸缩的方式进行处理，依靠横向扩展，通过不断增加廉价的商用服务器来提高计算和存储能力。用户可以轻松地在上面开发和运行处理海量数据的应用程序。Hadoop 主要有以下几个优点。

(1) 高可靠性：按位存储和处理数据的能力值得人们信赖。

(2) 高扩展性：在可用的计算机集群中分配数据并完成计算任务，这些集群可以方便地扩展到数以千计的节点规模。

(3) 高效性：能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

(4) 高容错性：能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

Hadoop 平台主要面向离线批处理应用，它是通过调度批量任务操作大规模的静态数据，因此，计算过程相对缓慢，有的查询可能会花几小时甚至更长时间才有结果，对于实时性要求更高的应用和服务则显得力不从心。MapReduce 是一种很好的集群并行编程模型，能够满足大部分应用的需求。虽然 MapReduce 是分布式/并行计算方面一个很好的模型，但它并不一定适合解决计算领域的任何问题。例如，那些需要实时获取计算结果的应用，像基于流量的点击付费模式的广告投放，基于实时用户行为数据分析的社交推荐，基于网页检索和点击流量的反作弊统计，等等。对于这些实时应用，MapReduce 并不能提供高效处理，因为处理这些应用逻辑需要执行多轮作业，或者需要将输入数据的粒度切分到很小。

现在也有一些基于 Hadoop 的系统来处理流式数据的系统，一般有以下几种方式。但它们也只是在一定程度上降低延时，总的任务调度模式仍属于批处理。

(1) 微型批处理 MapReduce：就是把流式的数据按照时间或者大小形成小的静态数据，然后分别启动 MapReduce 来计算。这种方式的缺点在于其延迟与数据片段的长度，以及分隔片段、初始化处理任务的附加开销成正比。小的分段会降低延迟，增加附加开销，并且使分段间的依赖管理更加复杂(例如一个分段可能会需要前一个分段的信息)。反之，大的分段会增加延迟。最优化的分段大小依赖于具体的应用。

(2) 连续的 MapReduce：像 Hadoop Online 这样的系统，通过作业内的数据传输流水线和作业间的数据传输流水线机制，实现了在线聚合和连续查询。当前 MapReduce 模型中，只有 Map 中间结果完全产生后，Reduce 才会过来拖数据，等所有 Map 数据都拖成功后，才能计算。Hadoop Online 实现了 Map 到 Reduce 间的数据流水线，使得 Map 在产生部分数据后，就可以送到 Reduce 端，以便 Reduce 可以提前或者定期计算。

(3) 动态添加输入：百度的一种实现，用来解决计算时数据还没有到位的问题。作业可以在数据还没有完全到位的情况下启动，当新数据累积到一定量时，通过一个命令行接口，向运行中的作业动态增加新的输入。这种方式大大减少了处理大数据作业时等待数据到位的时间，在依次执行多个作业时，也会有时间收益。

这类基于 MapReduce 进行流式处理的方案有三个主要缺点。

(1) 将输入数据分割成固定大小的片段，再由 MapReduce 平台处理，缺点在于处理延

迟与数据片段的长度、初始化处理任务的开销成正比。小的分段会降低延迟，增加附加开销，而且分段之间的依赖管理更加复杂（例如一个分段可能会需要前一个分段的信息）；反之，大的分段会增加延迟。最优化的分段大小取决于具体应用。

（2）为了支持流式处理，MapReduce 需要被改造成 Pipeline 的模式，而不是 Reduce 直接输出。考虑到效率，中间结果最好只保存在内存中，等等。这些改动使得原有的 MapReduce 框架的复杂度大大增加，不利于系统的维护和扩展。

（3）用户被迫使用 MapReduce 的接口来定义流式作业，这使得用户程序的可伸缩性降低。

除了 MapReduce 计算模型之外，以 Swift[①] 为代表的工作流计算模式，以 Pregel 为代表的图计算模式，也都可以处理包含大规模计算任务的应用流程和图算法。Swift 系统作为科学工作流和并行计算之间的桥梁，是一个面向大规模科学和工程工作流的快速、可靠的定义、执行和管理的并行化编程工具。Swift 采用结构化的方法管理工作流的定义、调度和执行，它包含简单的脚本语言 SwiftScript，SwiftScript 可以用来简洁地描述基于数据集类型和迭代的复杂并行计算，同时还可以对不同数据格式的大规模数据进行动态的数据集映射。运行时系统提供一个高效的工作流引擎用来进行调度和保证负载均衡，它还可以与 PBS 和 Condor 等资源管理系统进行交互，完成任务。

Pregel 是一种面向图算法的分布式编程框架，可以用于图遍历、最短路径、PageRank 计算等。它采用迭代的计算模型是，在每一轮，每个顶点处理上一轮收到的消息，并给其他顶点发出消息，并更新自身状态和拓扑结构（出、入边）等。

3.4.2 实时交互计算

当今的实时计算一般都需要处理海量数据，除了要满足非实时计算的一些需求（如计算结果准确）以外，还需要能够实时响应计算结果，一般实时响应时间的要求为秒级。实时计算一般可以分为以下两种应用场景。

（1）数据量巨大且不能提前计算出结果，但要求对用户的响应时间是实时的。该种情形主要用于特定场合下的数据分析处理。当数据量庞大，同时发现无法穷举所有可能条件的查询组合，或者大量穷举出来的条件组合无用的时候，实时计算就可以发挥作用。即将计算过程推迟到查询阶段进行，但需要为用户提供实时响应。在这种情形下，也可以将一部分数据提前处理，再结合实时计算结果，以提高处理效率。

（2）数据源是实时的不间断的，要求对用户的响应时间也是实时的。数据源实时不间断的也称为流式数据。所谓流式数据是指将数据看作是数据流的形式来处理。数据流是在时间分布和数量上无限的一系列数据记录的集合体，数据记录是数据流的最小组成单元。例如，在物联网领域传感器产生的数据可能是源源不断的。对于流式处理系统我们将分开在 3.4.3 节具体介绍。实时的数据计算和分析可以动态地对数据进行分析统计，对于系统的状态监控、调度管理具有重要的实际意义。

海量数据的实时计算过程可以被划分为以下三个阶段：数据的产生与收集阶段、传输与分析处理阶段、存储和对外提供服务阶段，如图 3-8 所示。

① http://www.ci.uchicago.edu/Swift/main/



图 3-8 实时计算过程

数据实时采集在功能上需要保证可以完整地收集到所有数据，为实时应用提供实时数据；响应时间上要保证实时性、低延迟性；配置简单，部署容易；系统稳定可靠等。目前，互联网企业的海量数据采集工具前面介绍过，包括 Facebook 开源的 Scribe、LinkedIn 开源的 Kafka、Cloudera 开源的 Flume、淘宝开源的 TimeTunnel、Hadoop 的 Chukwa 等，这些工具均可以满足每秒数百 MB 的日志数据采集和传输需求。

数据实时计算是这样的。传统的数据操作，首先将数据采集并存储在数据库管理系统(DBMS)中，然后通过 query 和 DBMS 进行交互，得到用户想要的答案。整个过程中，用户是主动的，而 DBMS 是被动的。但是，对于现在大量存在的实时数据，它们的实时性强，数据量大，数据格式多种多样，传统的关系型数据库架构并不合适。新型的实时计算架构一般都是采用海量并行处理 MPP 的分布式架构，数据的存储及处理会分配到大规模的节点上进行，以满足实时性的要求。在数据的存储上，则采用大规模分布式文件系统，比如 Hadoop 的 HDFS 文件系统，或是新型的 NoSQL 分布式数据库。

实时查询服务的实现可以分为三种方式：①全内存，直接提供数据读取服务，定期 dump 到磁盘或数据库进行持久化；②半内存，使用 Redis①、Memcache②、MongoDB③、BerkeleyDB 等数据库提供数据实时查询服务，由这些系统进行持久化操作；③全磁盘，使用 HBase 等以分布式文件系统(HDFS)为基础的 NoSQL 数据库，而 key-value 引擎的关键是设计好 key 的分布。

实时和交互式计算技术中，Google 的 Dremel 系统表现最为突出。Dremel 是 Google 的“交互式”数据分析系统。可以组建成规模上千的集群，处理 PB 级别的数据。作为 MapReduce 的发起人，Google 开发了 Dremel 系统，将处理时间缩短到秒级，作为 MapReduce 的有力补充。Dremel 作为 Google BigQuery 的 report 引擎，是一个很大的成功。和 MapReduce 一样，Dremel 也需要和数据运行在一起，将计算移动到数据上面。它需要 GFS 这样的文件系统作为存储层。Dremel 支持一个嵌套的数据模型，类似于 JSON。而传统的关系模型，由于不可避免的有大量的 Join 操作，在处理如此大规模的数据的时候，往往是有心无力。Dremel 同时还使用列式存储，分析的时候，可以只扫描需要的那部分数据，减少 CPU 和磁盘的访问量。同时列式存储是压缩友好的，使用压缩，可以减少存储量，以便发挥最大的效能。

Spark 是由加州大学伯克利分校 AMP 实验室开发的实时数据分析系统，它采用一种与 Hadoop 相似的开源集群计算环境，但是 Spark 在任务调度、工作负载优化等方面的设计和表现更加优越。Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。Spark 是利用 Scala 语言实现的，它将 Scala 用作其应用程序框架。Spark

① http://redis.io/
② http://memcached.org/
③ https://www.mongodb.org/

和 Scala 能够紧密集成，其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集。创建 Spark 可以支持分布式数据集上的迭代作业，而且支持对数据的快速统计分析，是对 Hadoop 的有效补充。它也可以在 Hadoop 文件系统中并行运行，通过名为 Mesos 的第三方集群框架支持此功能。Spark 可用来构建大型的、低延迟性的数据分析应用程序。

由 Cloudera 公司最近发布的 Impala 系统，类似于 Google 的 Dremel 系统，是一个有效的大数据实时查询工具。Impala 能在 HDFS 或 HBase 上提供快速、交互式 SQL 查询，它除了使用统一的存储平台之外，还使用了与 Hive 相同的 Metastore 及 SQL 语法等，为批处理和实时查询提供了一个统一的平台。

3.4.3 流计算

在很多实时应用场景中，比如实时交易系统、实时诈骗分析、实时广告推送、实时监控、社交网络实时分析等，数据量大，实时性要求高，而且数据源是实时不间断的。新到的数据必须马上处理完，不然后续的数据就会堆积起来，永远也处理不完。反应时间通常要求在秒级以下，甚至是毫秒级，这就需要一个高度可扩展的流式计算解决方案。

流计算就是为实时连续的数据类型而准备的。在数据不断变化的运动过程中实时地进行分析，捕捉到可能对用户有用的信息，并把结果发送出去。在整个过程中，数据分析处理系统是主动的，而用户却是处于被动接收的状态，如图 3-9 所示。

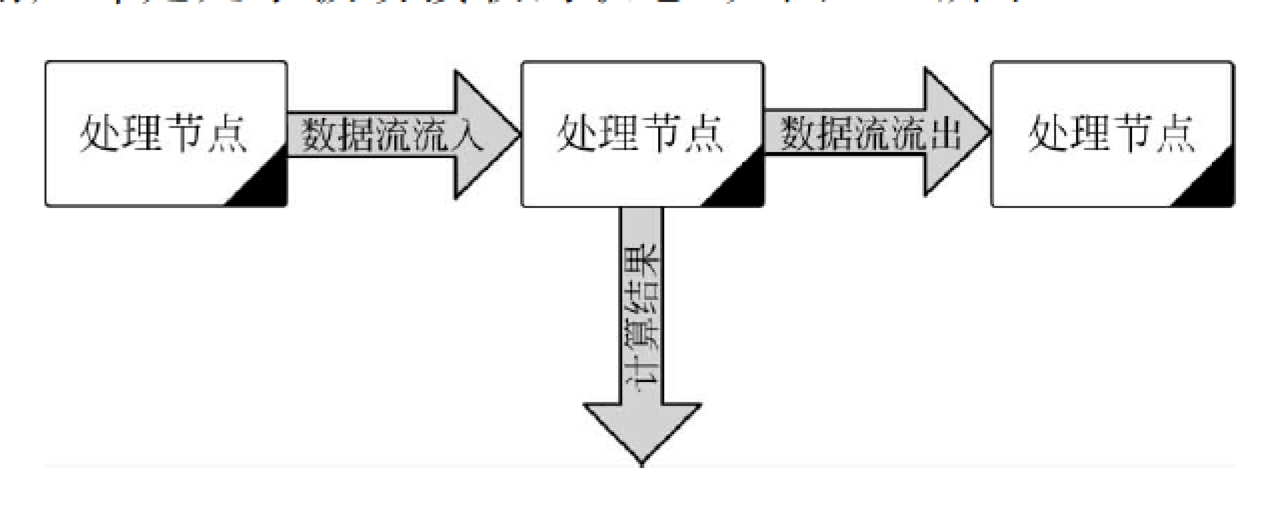


图 3-9 流计算过程

传统的流式计算系统，一般是基于事件机制，所处理的数据量也不大。新型的流处理技术，如 Yahoo 的 S4，主要解决的是高数据率和大数据量的流式处理。

S4 是一个通用的、分布式的、可扩展的、部分容错的、可插拔的平台。开发者可以很容易地在其上开发面向外界不间断流数据处理的应用。数据事件被分类路由到处理单元(Processing Elements，PE)，处理单元分析这些事件，并做如下的处理。

(1) 发出一个或多个可能被其他 PE 处理的事件；

(2) 发布结果。

S4 的设计主要由大规模应用在生产环境中的数据采集和机器学习所驱动。其主要特点如下。

(1) 提供一种简单的编程接口来处理数据流；

(2) 设计一个在普通硬件之上可扩展的高可用集群；

(3) 在每个处理节点使用本地内存，避免磁盘 I/O 瓶颈达到最小；

(4) 使用一个去中心的，对等架构，所有节点提供相同的功能和职责，没有担负特殊责任的中心节点，这大大简化了部署和维护；

(5) 使用可插拔的架构，使设计尽可能的既通用又可定制化；

(6) 友好的设计理念,易于编程,具有灵活的弹性。

S4 的设计和 IBM 的流处理核心 SPC 中间件有很多相同的特性。两个系统都是为了大数据量设计的,都具有能够使用用户定义的操作在持续数据流上采集信息的能力。两者主要的区别在架构的设计上,SPC 的设计源于 Publish/Subscribe 模式,而 S4 的设计是 MapReduce 和 Actor 模式的结合。Yahoo 相信因为其对等的结构,S4 的设计非常简单。集群中的所有节点都是等同的,没有中心控制。

SPC 是一种分布式的流处理中间件,用于支持从大规模的数据流中抽取信息的应用。SPC 包含为实现分布式的、动态的、可扩展的应用而需要的编程模式和开发环境,其编程模式包括用于申明和创建处理单元(PE)的 API,以及组装、测试、调试和部署应用的工具集。与其他流处理中间件不同的是,SPC 除了支持关系型的操作符外,还支持非关系型的操作符和用户自定义函数。

Storm[①] 是 Twitter 开发的一个类似于 Hadoop 的实时数据处理框架,这种高可拓展性,使得能处理高频数据和大规模的实时流数据计算解决方案应用于实时搜索、高频交易和社交网络上。Storm 有以下三大作用领域。

(1) 信息流处理(Stream Processing)。Storm 可以用来实时处理新数据和更新数据库,兼具容错性和可扩展性。

(2) 连续计算(Continuous Computation)。Storm 可以进行连续查询并把结果即时反馈给客户,比如将 Twitter 上的热门话题发送到客户端。

(3) 分布式远程过程调用(Distributed RPC)。Storm 可以用来并行处理密集查询,Storm 的拓扑结构是一个等待调用信息的分布函数,当它收到一条调用信息后,会对查询内容进行计算,并返回查询结果。

一个 Storm 集群和 Hadoop 集群表面上看很类似,但是 Hadoop 上运行的是 MapReduce 任务,而在 Storm 上运行的是拓扑,一个拓扑实际上定义的是一个消息流的处理的过程,简单来说,就是从一些数据源(叫做 Spout)产生的消息流,经过一些处理单元(叫做 Bolt)加工后产生新的消息流,这些消息流又接着被另外的加工单元处理,再产生其他的消息流。这些数据源(Spouts)和加工单元(Bolts)所组成的整个处理架构就是一个拓扑。

消息源 Spout 是 Storm 里面一个 Topology 里面的消息生产者,如图 3-10 所示。一般来说消息源会从一个外部源读取数据并且向 Topology 里面发出消息,消息源可以发射多条消息流 Stream。所有的消息处理逻辑被封装在 Bolts 里面。Bolts 可以做很多事情:过滤,聚合,查询数据库等。Bolts 可以简单地做消息流的传递。复杂的消息流处理往往需要很多步骤,从而也就需要经过很多 Bolts。一个 Bolt 也可以继续发射出多条消息流,被其他 Bolts 继续处理。

比如需要设计一个 Topology,来对一个句子里的单词进行词频统计,那么整个 Topology 看起来如图 3-11 所示。其中包含一个 Spout,用来从 Kestrel 队列中读取一个句子,把它输出成一个消息,发送给第一个 Bolt,进行单词切分。然后第二个 Bolt 汇总每个单词出现的次数,这样整体上一个 Spout 加上两个 Bolts 就构成了一个单词词频统计的拓扑。

① http://storm-project.net/

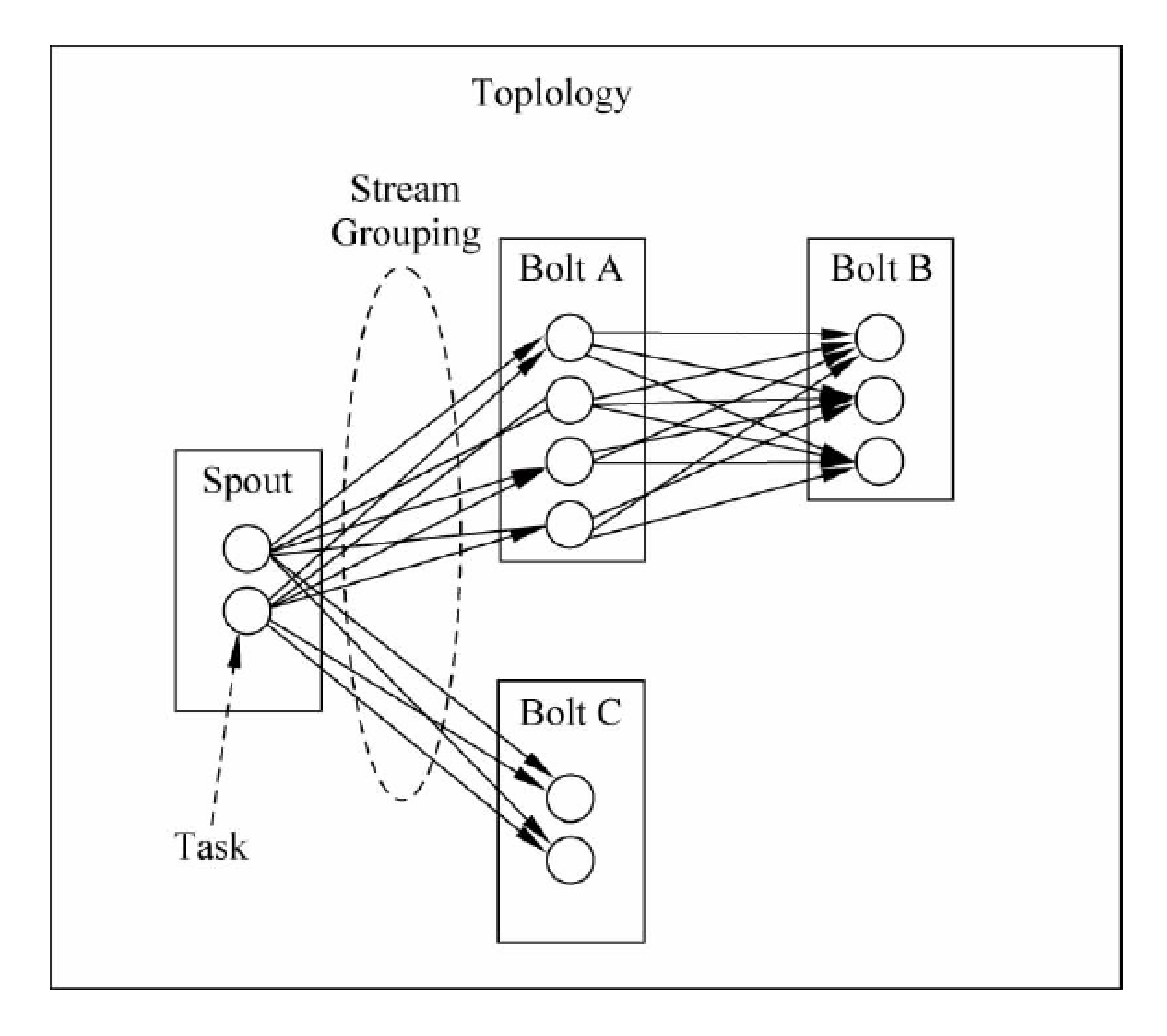


图 3-10 Storm 的 Topology 由 Spouts 和 Bolts 组成

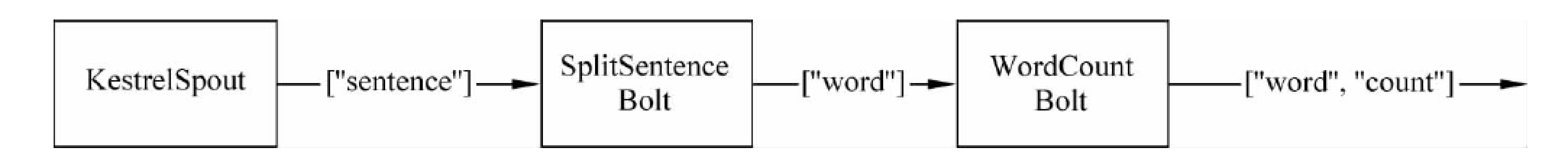


图 3-11 用于词频统计的 Topology 示例

3.5 数据交互展示

计算结果需要以简单直观的方式展现出来，才能最终为用户所理解和使用，形成有效的统计、分析、预测及决策，应用到生产实践和企业运营中，因此大数据的展现技术，以及数据的交互技术在大数据全局中也占据重要的位置。

Excel 形式的表格和图形化展示方式是人们熟知和使用已久的数据展示方式，也为日常的简单数据应用提供了极大的方便。华尔街的很多交易员还都依赖 Excel 和他们很多年积累和总结出来的公式来进行大宗的股票交易，而微软公司和一些创业者也看到市场潜力，在开发以 Excel 为展示和交互方式、结合 Hadoop 等技术的大数据处理平台。

人脑对图形的理解和处理速度大大高于文字。因此，通过视觉化呈现数据，可以深入展现数据中的潜在的或复杂的模式和关系。随着大数据的兴起，也涌现了很多新型的数据展现和交互方式，以及专注于这方面的一些创业公司。这些新型方式包括交互式图表，可以在网页上呈现，并支持交互，可以操作、控制图标，进行动画演示。另外，交互式地图应用如 Google 地图，可以动态标记、生成路线、叠加全景航拍图等，由于其开放的 API 接口，可以与很多用户地图和基于位置的服务应用结合，因而获得了广泛的应用。Google Chart Tools 也给网站数据可视化提供了很多种灵活的方式。从简单的线图、Geo 图、Gauges（测量仪），到复杂的树图，Google Chart Tools 提供了大量设计优良的图表工具。

大数据时代也诞生了很多新兴的大数据可视化技术及相应的创业公司，能够将数据所蕴含的信息与可视化展示有机地结合起来的“信息图”方式，目前大行其道。诞生于斯坦福

大学的大数据创业公司 Tableau 能够将数据运算与美观的图表完美地结合在一起。Tableau 的设计与实现理念是：界面上的数据越容易操控，公司对自己所在业务领域里的所作所为到底是正确还是错误，就能了解得越透彻。快速处理，便捷共享，是 Tableau 的另一大特性，这将使得用户使用数据的积极性大大增加。另一家大数据可视化创业公司 Visually 以丰富的信息图资源而著称，它是一个社会化的信息图创作分享平台。很多用户乐意把自己制作的信息图上传到网站中与他人分享，信息图极大地刺激视觉表现，促进用户间相互学习、讨论。

此外，3D 数字化渲染技术也被广泛地应用在很多领域，如数字城市、数字园区、模拟与仿真、设计制造等，具备很高的直观操作性。现代的虚拟现实 VR 和增强现实 AR 技术通过计算机技术，将虚拟的信息应用到真实世界，真实的环境和虚拟的物体实时地叠加到同一个画面或空间同时存在。结合虚拟 3D 的数字模型和真实生活中的场景，提供了更好的现场感和互动性。通过 VR/AR 技术，用户可以和虚拟的物体进行交互，如试戴虚拟眼镜、试穿虚拟衣服、驾驶模拟飞行器等。在德国，工程技术人员在进行机械安装、维修、调试时，通过头盔显示器，可以将原来不能呈现的机器内部结构，以及它的相关信息、数据完全呈现出来。

现代的体感技术，如微软的 Kinect 以及 Leap 公司的 Leap Motion 体感控制器，能够检测和感知到人体的动作及手势，进而将动作转化为对计算机及系统的控制，使人们摆脱了键盘、鼠标、遥控器等传统交互设备的束缚，直接用身体和手势来与计算机和数据交互。当今热门的可穿戴式技术，如 Google 眼镜，则有机地结合了大数据技术、增强现实及体感技术。随着数据的完善和技术的成熟，我们可以实时地感知我们周围的现实环境，并且通过大数据搜索、计算，实现对周围的建筑、商家、人群、物体的实时识别和数据获取，并叠加投射在我们的视网膜上，这样可以实时地帮助我们工作、购物、休闲等，提供极大的便利。

3.5.1　数据可视化基础

数据可视化主要旨在借助于图形化手段，清晰有效地传达与沟通信息。但是，这并不就意味着，数据可视化就一定因为要实现其功能用途而令人感到枯燥乏味，或者是为了看上去绚丽多彩而显得极端复杂。为了有效地传达思想观念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，从而实现对于相当稀疏而又复杂的数据集的深入洞察。然而，设计人员往往并不能很好地把握设计与功能之间的平衡，从而创造出华而不实的数据可视化形式，无法达到其主要目的，也就是传达与沟通信息。

数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。当前，在研究、教学和开发领域，数据可视化乃是一个极为活跃而又关键的方面。“数据可视化”这条术语实现了成熟的科学可视化领域与较年轻的信息可视化领域的统一。

数据可视化领域的起源可以追溯到 20 世纪 50 年代计算机图形学的早期。当时，人们利用计算机创建出了首批图形图表。1987 年，由布鲁斯·麦考梅克、汤姆斯·蒂凡提和玛克辛·布朗所编写的美国国家科学基金会报告 *Visualization in Scientific Computing*（意为“科学计算之中的可视化”），对于这一领域产生了大幅度的促进和刺激。这份报告之中强调了新的基于计算机的可视化技术方法的必要性。随着计算机运算能力的迅速提升，人们

创建了规模越来越大，复杂程度越来越高的数值模型，从而造就了形形色色体积庞大的数值型数据集。同时，人们不但利用医学扫描仪和显微镜之类的数据采集设备产生大型的数据集，而且还利用可以保存文本、数值和多媒体信息的大型数据库来收集数据。因而，就需要高级的计算机图形学技术与方法来处理和可视化这些规模庞大的数据集。

短语“Visualization in Scientific Computing”（意为“科学计算之中的可视化”）后来变成了“Scientific Visualization”（即“科学可视化”），

而前者最初指的是作为科学计算之组成部分的可视化，也就是科学与工程实践当中对于计算机建模和模拟的运用。更近一些的时候，可视化也日益尤为关注数据，包括那些来自商业、财务、行政管理、数字媒体等方面的大型异质性数据集合。20 世纪 90 年代初期，人们发起了一个新的称为“信息可视化”的研究领域，旨在为许多应用领域之中对于抽象的异质性数据集的分析工作提供支持。因此，目前人们正在逐渐接受这个同时涵盖科学可视化与信息可视化领域的新生术语“数据可视化”。

自那时起，数据可视化就是一个处于不断演变之中的概念，其边界在不断地扩大；因而，最好是对其加以宽泛的定义。数据可视化指的是技术上较为高级的技术方法，而这些技术方法允许利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。与立体建模之类的特殊技术方法相比，数据可视化所涵盖的技术方法要广泛得多。

3.5.2 数据可视化模式

数据可视化分为科学可视化和信息可视化这两个主要分支。

科学可视化，处理科学数据，面向科学和工程领域的科学可视化，研究带有空间坐标和几何信息的三维空间测量数据、计算模拟数据和医疗影像数据等，重点探索如何有效地呈现数据中几何、拓扑和形状特征。信息可视化，处理对象是非结构化、非几何的抽象数据，如金融交易、社交网络和文本数据，其核心挑战是如何针对大尺度高维数据减少视觉混淆对有用信息的干扰。

1. 科学可视化

面向的领域主要是自然科学，如物理、化学、气象气候、航空航天、医学、生物学等各个学科，这些学科需要对数据和模型进行解释、操作与处理，旨在寻找其中的模式、特点、关系以及异常情况。

标量场可视化。标量指单个数值，即在每个记录的数据点上有一个单一的值，标量场指二维、三维或四维空间中每个采样处都有一个标量值的数据场。可视化数据场 $f(x,y,z)$ 的标准做法有三种：①将数值直接映为颜色或透明度，如用颜色表达地球表面的温度分布；②根据需要抽取并连接满足 $f(x,y,z)=c$ 的点集，并连接为线或面，称为等值线或等值面方法，如地图中的等高线，标准的算法有移动四边形或移动立方体；③将三维标量数据场看成能产生、传输和吸收光的媒介，光源透过数据场后形成半透明影像，称为直接体绘制方法，这种方法可以以透明层叠的方式显示内部结构，为观察三维数据场全貌提供了极好的交互浏览工具。

向量场可视化。向量场在每个采样点处都是一个向量（一维数据组）。向量代表某个方向或趋势，例如风向等。向量场可视化主要关注点是其中蕴含的流体模式和关键特征区域。

在实际应用中,由于二维或三维流场是最常见的向量场,所以流场可视化是向量场可视化中最重要的组成部分。除了通过拓扑或几何方法计算向量场的特征点、特征线或特征区域外,对向量场直接进行可视化的方法包括以下三类。

(1) 粒子对流法,其关键思想是模拟粒子在向量场中以某种方式流动,获得的几何轨迹可以反映向量场的流体模式。这类方法包括流线、流面、流体、迹线和脉线等。

(2) 将向量场转换为一帧或多帧的纹理图像,为观察者提供直观的影像展示。标准的做法有随机噪声纹理、线积分卷积(LIC)等。

(3) 采用简化易懂的图标编码单个或简化后的向量信息,可提供详细信息的查询与计算,标准做法有线条、箭头和方向标志符等。

张量场可视化。方法分为基于纹理、几何和拓扑三类。基于纹理的方法,将张量场转换为静态图像或动态图像序列,图释张量场的全局属性,其思想是将张量场简化为向量场进而采用线性积分法、噪声纹理法等方法显示。基于几何的方法显示地生成刻画某类张量场的属性的几何表达,其中,图标法采用某种几何形式表达单个张量,如椭球和超二次曲面;超流线法将张量转换为向量,再沿主特征方向进行积分,形成流线、流面或流体。基于拓扑的方法计算张量场的拓扑特征(如关键点、奇点、灭点、分叉点和退化线等),依次将感兴趣区域部分分为具有相同属性的子区域,并建立对应的图结构,实现拓扑简化、拓扑跟踪和拓扑显示,基于拓扑的方法可以有效地生成多变量场的定性结构,快速构造全局流场结构,特别适合于数值模拟或实验模拟生成的大尺度数据。

2. 信息可视化

信息可视化处理的对象是抽象的、非结构化数据集(如文本、图表、层次结构、地图、软件、复杂系统等)。与科学可视化相比,信息可视化更关注抽象、高维数据。此类数据通常不具有空间中位置的属性,因此要根据特定数据分析的需求,决定数据元素在空间的布局。因为信息可视化的方法与所针对的数据类型紧密相关,所以通常按数据类型分为如下几类。

(1) 时空数据可视分析。

(2) 层次与网络结构数据可视化。

(3) 文本和跨媒体数据可视化。

(4) 多变量数据可视化。

3.5.3 数据可视化工具

1. 可视化基础工具

(1) D3.js。D3是一个用动态图形显示数据的JavaScript库,一个数据可视化的工具,如图3-12所示。

D3兼容W3C标准,并且利用广泛实现的SVG、JavaScript和CSS标准。它是早期的Protovis框架的继承者。与其他的类库相比,D3对视图结果有很大的可控性。

D3可以让用户随心所欲地把数据绑定到一个文档对象模型(DOM),然后应用数据驱动转换到文档中。例如,可以使用D3从数字数组生成一个HTML表。或者,使用相同的数据来创建平滑的过渡和互动的交互式SVG条形图。

D3不是一个整体框架,没有去尝试涵盖所有功能。相反,D3致力于解决如何基于数据

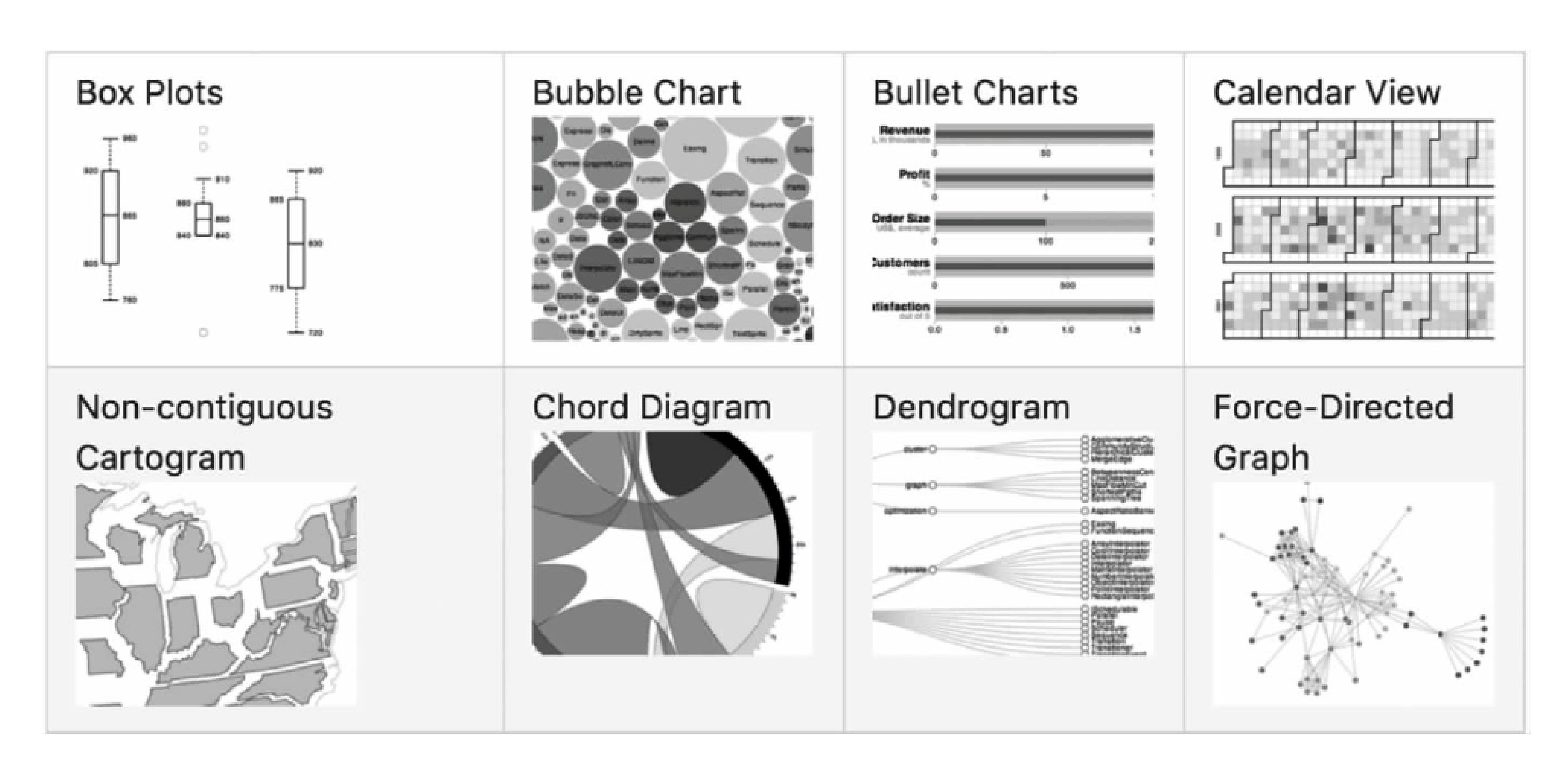


图 3-12　D3 可视化工具

的文档高效操作的问题，这使得 D3 具有非常高的灵活性。D3 非常快，而且开销很小。D3 支持大型数据的处理，提供动态交互。

(2) ECharts。ECharts 是一个纯 JavaScript 的图表库，可以流畅地运行在 PC 和移动设备上，兼容当前绝大部分浏览器(IE8/9/10/11，Chrome，Firefox，Safari 等)，底层依赖轻量级的 Canvas 类库 ZRender，提供直观、生动、可交互、可高度个性化定制的数据可视化图表，如图 3-13 所示。

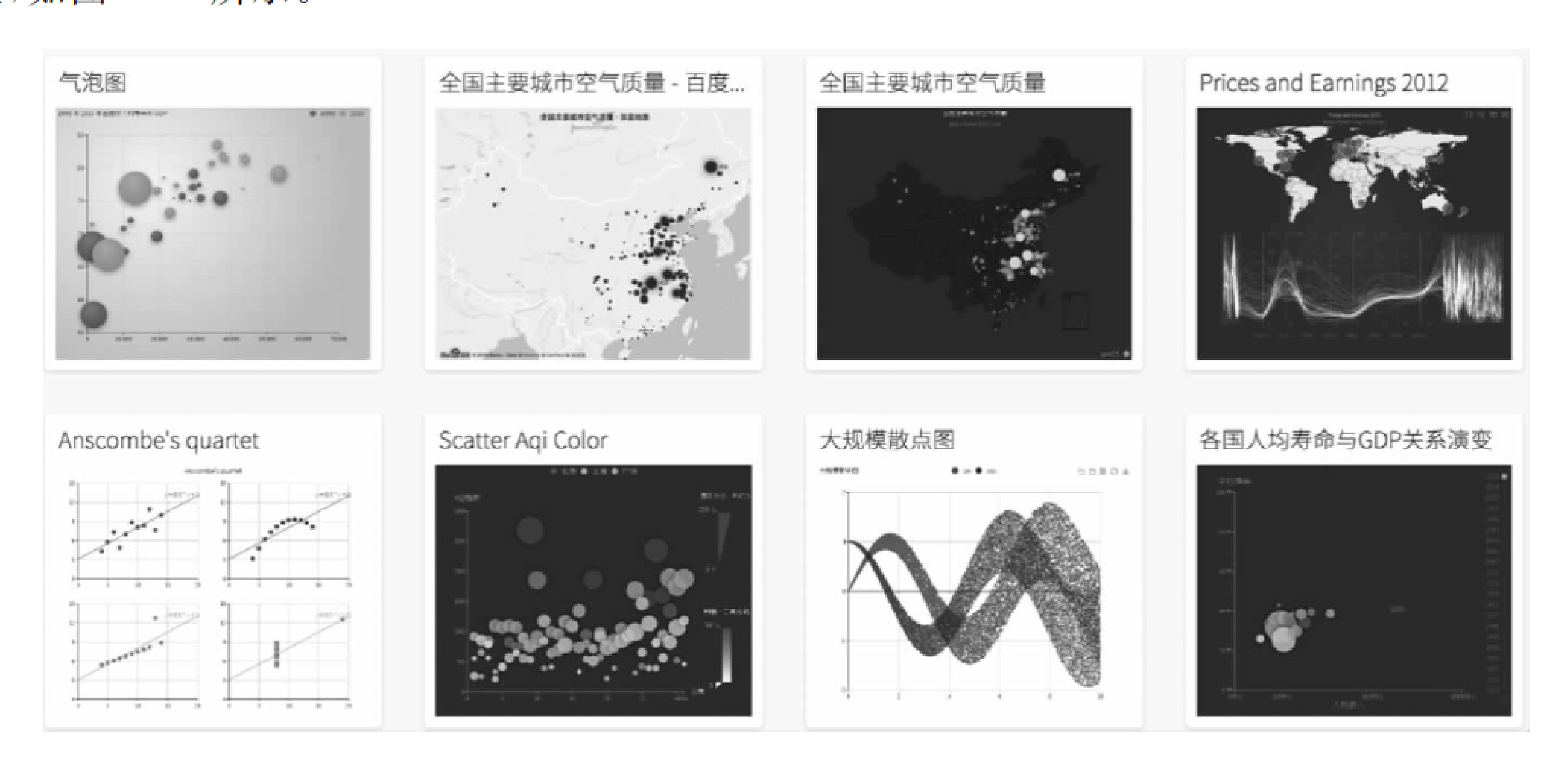


图 3-13　Echarts 图表库

ECharts 提供了常规的折线图、柱状图、散点图、饼图、K 线图，用于统计的盒形图，用于地理数据可视化的地图、热力图、线图，用于关系数据可视化的关系图、treemap，多维数据可视化的平行坐标，还有用于 BI 的漏斗图、仪表盘，并且支持图与图之间的混搭。

用户可以在下载界面下载包含所有图表的构建文件，如果只是需要其中一两个图表，又嫌包含所有图表的构件文件太大，也可以在在线构件中选择需要的图表类型后自定义构件。

ECharts 3 开始独立出了"坐标系"的概念，支持了直角坐标系(catesian，同 grid)、极坐

标系(polar)、地理坐标系(geo)。图表可以跨坐标系存在,例如,折、柱、散点等图可以放在直角坐标系上,也可以放在极坐标系上,甚至可以放在地理坐标系中。

同时,EChart 针对移动端进行了优化,提供深度的交互式探索,借助 Canvas 的能力,ECharts 在散点图中能够轻松展现上万甚至超过十万的数据。ECharts 3 开始加强了对多维数据的支持。除了加入了平行坐标等常见的多维数据可视化工具外,对于传统的散点图等,传入的数据也可以是多个维度的。配合视觉映射组件 visualMap 提供的丰富的视觉编码,能够将不同维度的数据映射到颜色、大小、透明度、明暗度等不同的视觉通道。ECharts 由数据驱动,数据的改变驱动图表展现的改变。因此动态数据的实现也变得异常简单,只需要获取数据,填入数据,ECharts 会找到两组数据之间的差异然后通过合适的动画去表现数据的变化。配合 timeline 组件能够在更高的时间维度上去表现数据的信息。

2. 商用可视化软件——Tableau

Tableau 公司将数据运算与美观的图表完美地嫁接在一起,如图 3-14 和图 3-15 所示。它的程序很容易上手,各公司可以用它将大量数据拖放到数字"画布"上,转眼间就能创建好各种图表。这一软件的理念是,界面上的数据越容易操控,公司对自己在所在业务领域里的所作所为到底是正确还是错误,就能了解得越透彻。

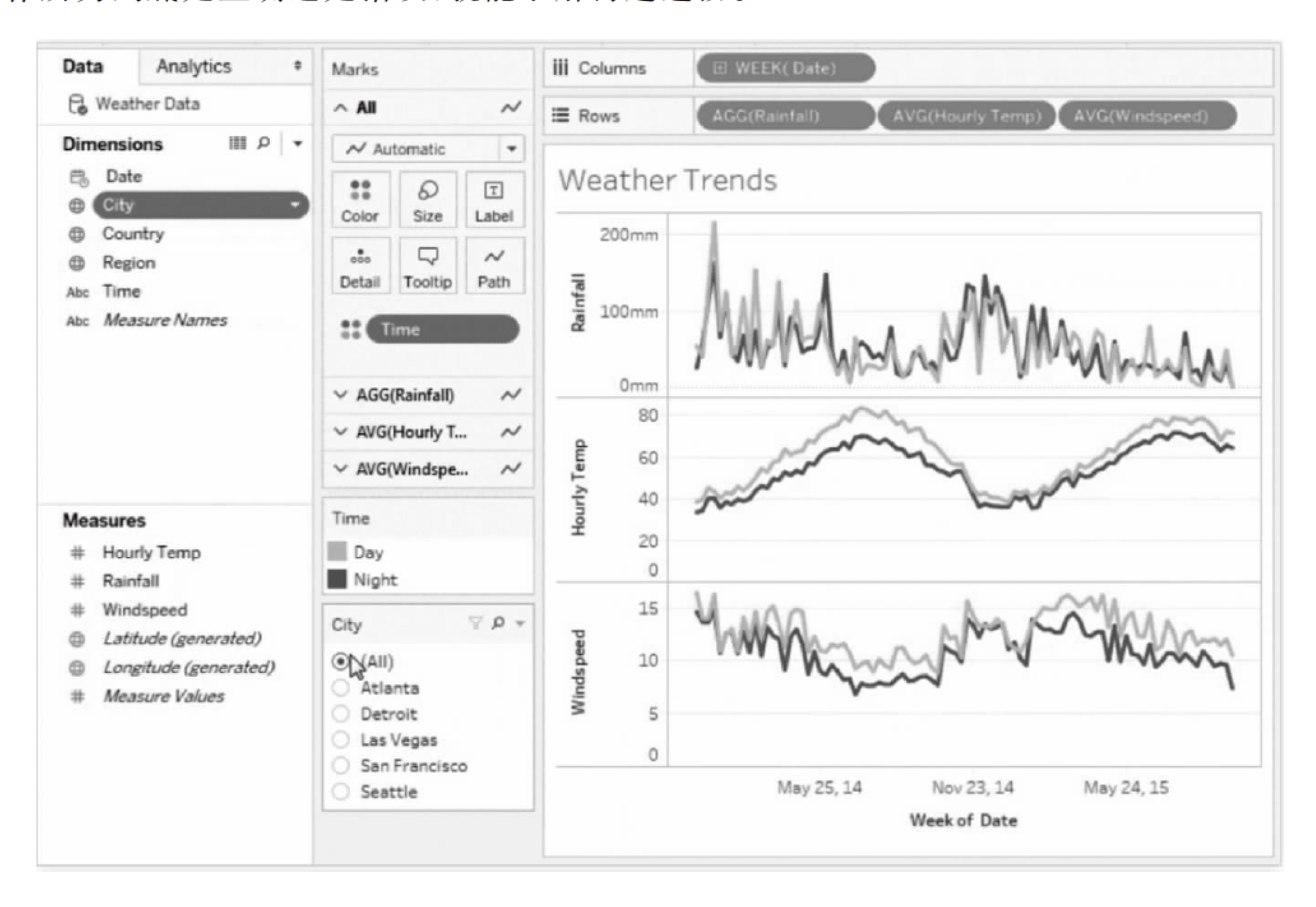


图 3-14　Tableau 可视化实例

Tableau 目前有 6 大软件产品:Tableau Desktop、Tableau Server、Tableau Online、Tableau Mobile、Tableau Public 以及 Tableau Reader。

Tableau 比现有解决方案快 10～100 倍。它根据人的思维方式设计,在画布上拖放,无论数据是位于电子表格中、SQL 数据库中、Hadoop 中还是在云端,都可以连接到任何数据;一键即可访问大数据;无须编写代码即可合并不同的数据源。

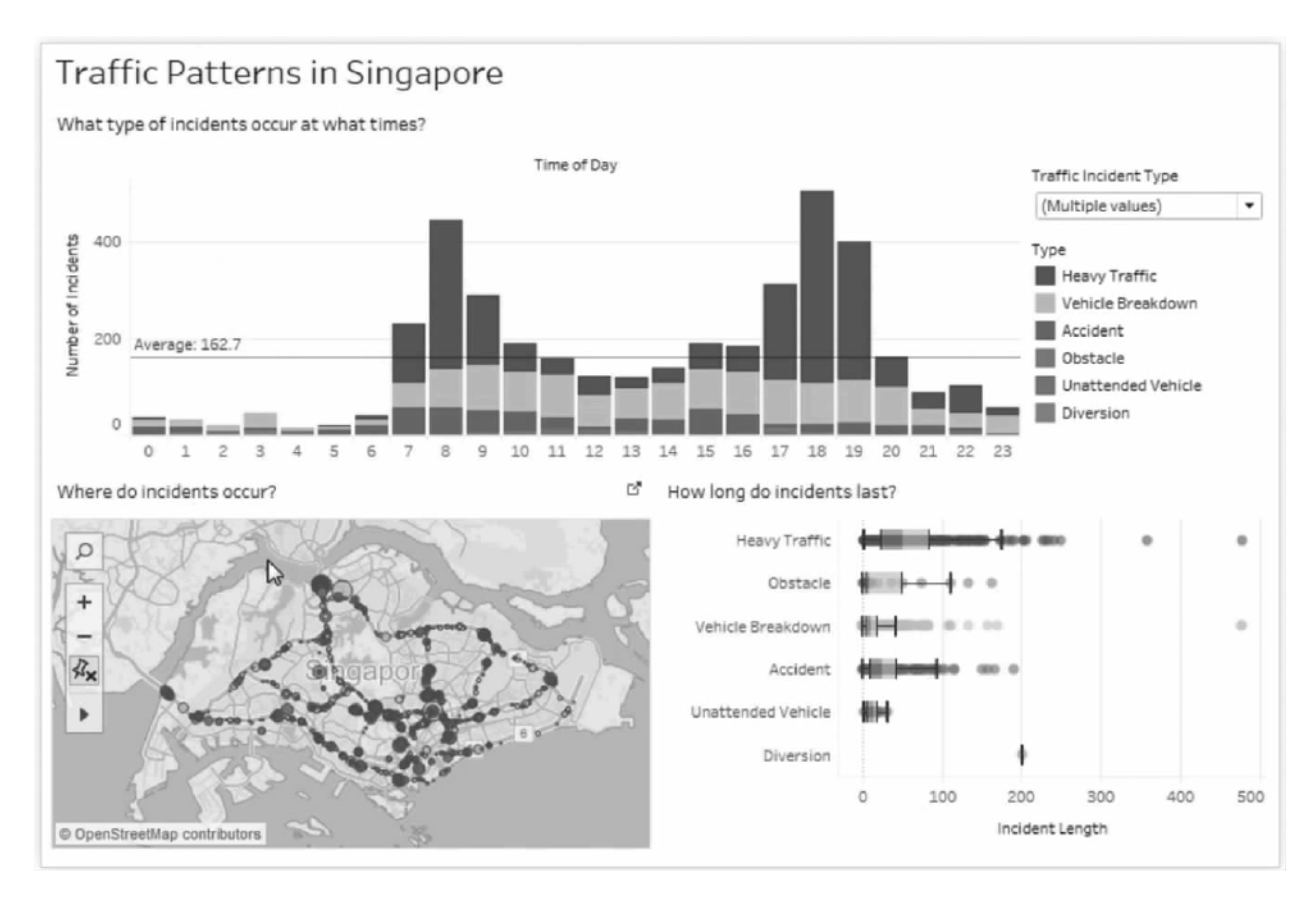


图 3-15 Tableau 可视化案例

3.6 大数据应用

大数据在过去几年得到了全社会的关注和快速的发展，几乎在每个行业都可以见到大数据应用的影子。大数据的应用范围越来越广，应用的行业也越来越多，我们几乎每天都可以看到大数据的一些新奇应用，大数据的价值也已经体现在方方面面。大数据目前较多的应用领域主要有互联网、金融、医疗、环保、工业制造、教育、政府等行业，应用的环境也不尽相同，具体应用场景介绍详见 2.3 节。如图 3-16 所示是大数据应用架构图。

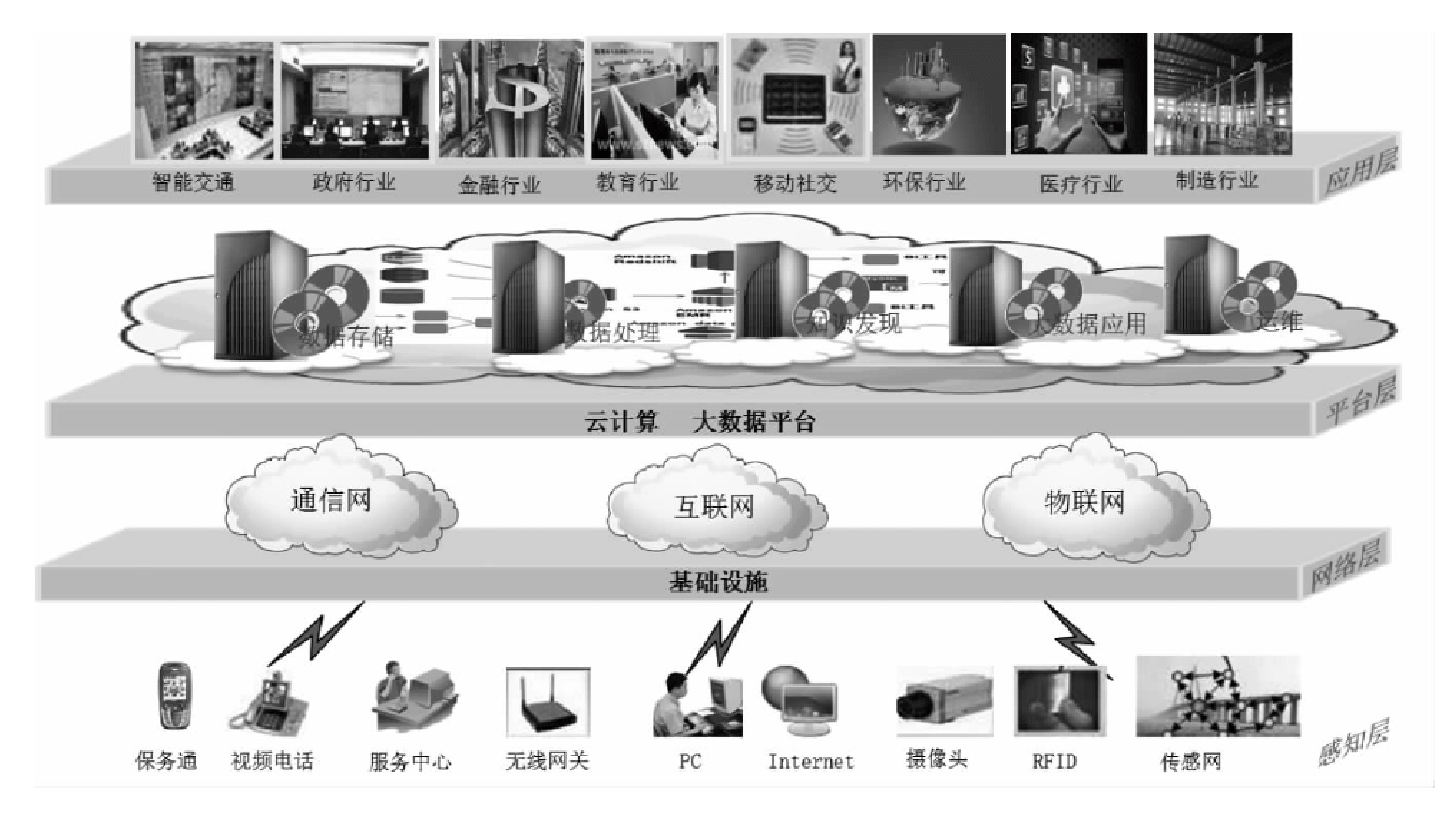


图 3-16 大数据应用架构图

3.7 运营管理

大数据平台在完成数据的采集、存储、处理、展示和应用之外，其自身的运营管理也非常重要，大数据平台的运营管理主要包括以下几个方面。

(1) 监控：对平台的硬件、软件、网络、服务及应用进行实时监控。

(2) 告警：基于监控信息及管理设置，当出现影响平台运营及服务的情况发生时，能够进行故障告警。

(3) 备份：对平台上的重要数据及状态进行备份。

(4) 恢复：基于备份信息进行数据及状态恢复。

(5) 优化：基于平台配置及平台的运行状况，对平台的软件、硬件、网络、服务及应用进行优化。

3.8 安全管理

大数据平台在给政府、企业、社会以及个人用户带来极大便利的同时，也促生了不同于以往的安全问题和威胁。在传统的安全防护体系中，“防火墙”起着至关重要的作用。防火墙是一种形象的说法，其实它是一种计算机硬件和软件的组合，在内部网络与外部网络之间建立起一个安全网关，从而保护内部网络免受外部非法用户的侵入。然而云计算架构很多时候是为多租户服务的，很多不同用户的应用都运行在同一云数据中心内部，这就打破了传统的安全体系中的内外之分。作为企业和用户来说，不仅要防范来自数据中心外部的攻击，还要提防云服务提供商，以及潜藏在云数据中心内部的其他别有用心的用户。同时云计算平台有大量的计算集群，如果被黑客控制，可以发动进行大规模的非法计算或大规模的攻击行为，比如利用这些服务器暴力破解政府重要部门的密码等。同时，大数据平台储存了大量有价值的信息，容易为不法分子所垂涎，一旦遭到入侵，损失巨大，对数据的安全保障也至关重要。

大数据平台的安全管理，需要从以下几方面着手。

(1) 物理安全。早期的也是最基础的安全涉及的是信息系统的物理安全，即整个系统所处的场所和环境的安全、设备和设施的安全，以及整个系统可靠运行等方面，这些是信息系统安全运行的基本保障。

从物理层面出发，系统物理安全技术应确保信息系统的安全性、保密性、可用性、完整性，比如门禁保安、机房建设、综合布线、通信线路的要求，机房应具备一定的防火防盗、温湿度控制能力、一定的应急供配电能力以保证系统的可用性；通过设备访问控制、边界保护、设备及网络资源管理等措施确保信息系统的保密性和完整性；通过容错、故障恢复、系统灾难备份等措施确保信息系统可用性。为保证系统整体的正常运行，还需要有设备备份、网络性能监测、设备运行状态监测、报警监测的要求。

(2) 网络安全。20 世纪 80 年代的信息系统，就做到了物理上的安全隔离和可靠运行，具备了基本的安全保障。然而到了 20 世纪 90 年代，随着网络的出现和发展，信息能够通过网络进行远程传输和交换，因而安全防护也就不再局限于信息系统的物理隔离，而是扩展到了整个网络可以到达的范围。网络安全是指网络系统的硬件、软件及其系统中的数据受到

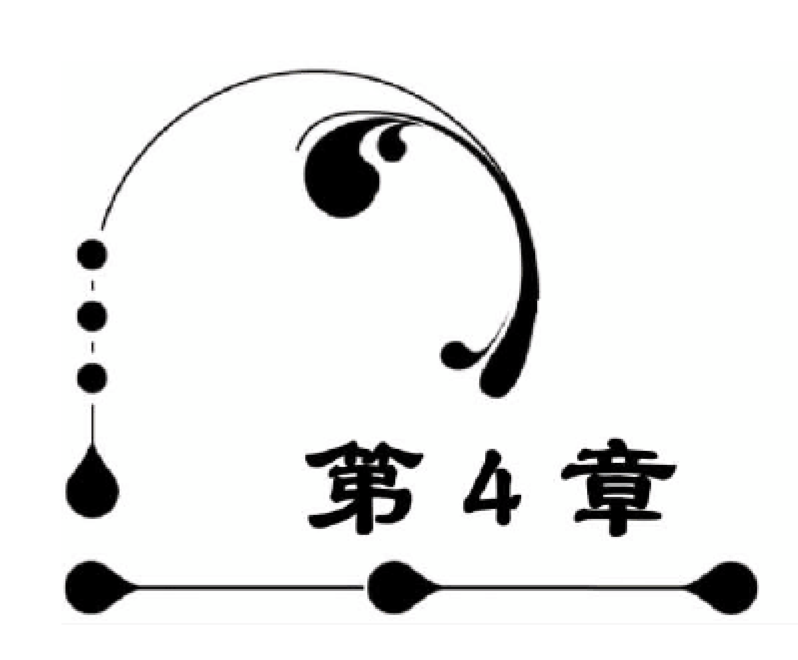
保护,不因偶然的或者恶意的原因而遭受到破坏、更改、泄漏,系统可以连续可靠正常地运行,网络服务不中断。网络安全包含网络设备安全、网络信息安全、网络软件安全。从广义来说,凡是涉及网络上信息的保密性、完整性、可用性、真实性和可控性的相关技术和理论都是网络安全的研究范畴。建立网络安全保护措施的目的是确保经过网络传输和交换的数据不会发生增加、修改、丢失和泄漏等问题。从网络运行和管理者角度说,希望本地网络信息的访问、读写等操作受到保护和控制,避免出现“陷门”、病毒、非法存取、拒绝服务和网络资源非法占用和非法控制等问题,制止和防御网络黑客的攻击。对安全保密部门来说,他们希望能够对非法的、有害的或涉及国家机密的信息进行过滤和防堵,避免机要信息泄漏,避免对社会产生危害,给国家造成巨大损失。从社会教育和意识形态角度来讲,网络上不健康的内容,会阻碍社会的稳定和人类的发展,必须对其进行控制。

(3) 应用安全。信息一般都是通过应用系统来存取,因此,应用系统的安全也是确保信息安全的一个重要部分。常见的应用有 Web 应用、数据库服务、电子商务等,首先需要确保这些应用的安全,才能保障它们所管理维护的信息的安全。Web 业务是开放的交互业务,其安全性也面临很大的挑战。这涉及身份鉴别,数据访问权限管理,保护服务器不被非法授权访问,保护浏览器不被恶意代码如病毒和木马等侵袭,保护网页不被非法篡改,防止 SQL 注入等。针对应用安全的常见的安全防护手段包括身份认证、访问控制、入侵防护、正确设置浏览器安全选项、定期漏洞扫描加固等。而带有支付功能的电子商务应用对于安全防护要求更高,因为它直接涉及用户的经济财产,尤其是当今的移动电子商务,泄漏风险非常高。在这方面,除了常规的网络安全和应用安全手段之外,还涉及密钥管理、数字证书、身份认证鉴权、电子支付手段等,对于黑客攻击、病毒及木马的防护也尤其重要。

(4) 数据安全。在当今大数据时代,数据安全就上升到了非常重要的地位,因为数据的体量大,价值高。数据安全,一是数据防丢失,主要是采用现代信息存储手段对数据进行主动防护,如磁盘阵列、数据备份和恢复、异地容灾等;二是数据防泄漏,首先可以采用现代密码算法对数据进行主动保护,如数据加密、数据完整性检查、双向强身份认证等;另外一方面需要防止数据被非法访问和盗取,在数据的传输和处理过程中对数据的防护也很重要。

除了以上几个方面之外,管理是信息安全中最重要的部分。安全意识不强、责权不明、安全管理制度不健全及缺乏可操作性等都会带来泄漏风险。事前对于安全防范不重视,缺乏严密的安全管理、防护制度及流程,当出现安全风险和威胁时(如遭受攻击或内部人员操作违规等),无法进行实时的检测、监控、报告与预警,在事故发生后,也不能提供追溯线索、采取补救措施、加强防范,必然会导致严重的后果和损失。

大数据平台是大数据管理的技术基础,也是有效地将大数据经过清洗、梳理、转换,再进行加工和深度利用,能够形成数据资产和产生数据价值的基础处理架构和工具。大数据技术平台与传统信息技术体系的区别在于它能够处理大数据的多源异构性、高通量、大容量、实时性等需求,采用的方法包括云计算、分布式存储、分布式计算、并行处理架构、海量批处理、实时流处理、数据挖掘算法及模型、人工智能、深度学习、虚拟现实/增强现实等一系列新一代信息处理架构,因而也能有效地突破传统信息系统的瓶颈,充分发掘和实现数据价值。



大数据的数据整合、交换与交易

大数据的特点在于其多样性，数据具备多种类别和结构，来自不同的数据源。把一个行业的全样本数据整合起来，或是把不同种类数据整合在一起，进行汇总和交叉关联分析，就能得到有价值的发现，实现大数据的行业创新应用。同时，传统数据分析的基础是获取到数据的所有权，所以其应用范围有限。而在大数据时代，数据整合需要打破行业条块分割和数据孤岛，整合尽可能多的数据源，因此数据的交换共享非常重要。一个城市的大数据建设涉及政务、医疗、教育、交通、城市管理、公共安全、应急指挥等几十个方面，首先就需要在这些行业实现数据整合和交换。在此之上，还可以把数据加工变成服务，数据变成服务以后，也可以方便地整合在一起，形成新的服务和价值。这种整合可以发生在政府部门之间，企业之间，行业之间，也可以在这些实体之间交叉，具体取决于数据是如何被使用的。由于数据本身就是资产，具备价值，因此基于数据之间，以及数据衍生服务之间还可以进行数据交易，在未来，数据交易可能是最赚钱的数据业务。

大数据产业自 2012 年美国发布大数据国家战略开始，迎来了蓬勃的发展，自 2015 年开始，则进入了高速增长期。据研究机构预测，未来 5 年，全球大数据市场将保持 31.7%的年复合增长率，中国大数据市场的年复合增长率将高达 51.4%，大数据产业正在成为新的经济增长点。大数据价值发掘和体现的前提和基础是数据的开放、流通，通过整合利用数据资源，才能激发数据的市场和创新活力。

举例来说，当前最热门的互联网金融行业，需要形成企业和个人的信用评估，就需要把金融企业自身的客户信息，和工商、税务、通信、消费、旅游，乃至用户的社交信息整合起来，才能形成对客户的完整描述和评估。在智慧旅游产业，经营酒店的业主则需要获得天气、交通、景区等信息，才能优化经营和服务。

大数据的数据来源广泛，应用需求和数据类型都不尽相同，从数据的来源及其基本的流向，我们总结其最基本的处理流程是一致的，如图 4-1 所示。

数据源的界定：找到我们分析和处理所需要的数据源，数据源可能包括结构化的数据、半结构化的数据、非结构化的数据等。

数据的抽取和整合：对广泛异构的数据源进行抽取和集成，结果按照一定的标准进行统一存储。在抽取和整合过程中需要注重数据的质量和可信度，形成对数据的元数据(模式等)的描述，并做一定的聚合和关联处理。

数据分析：利用合适的数据分析技术对存储的数据进行分析，从中提取有价值的知识和信息。主要的分析手段有辅助决策、商业智能(BI)、推荐、预测等。

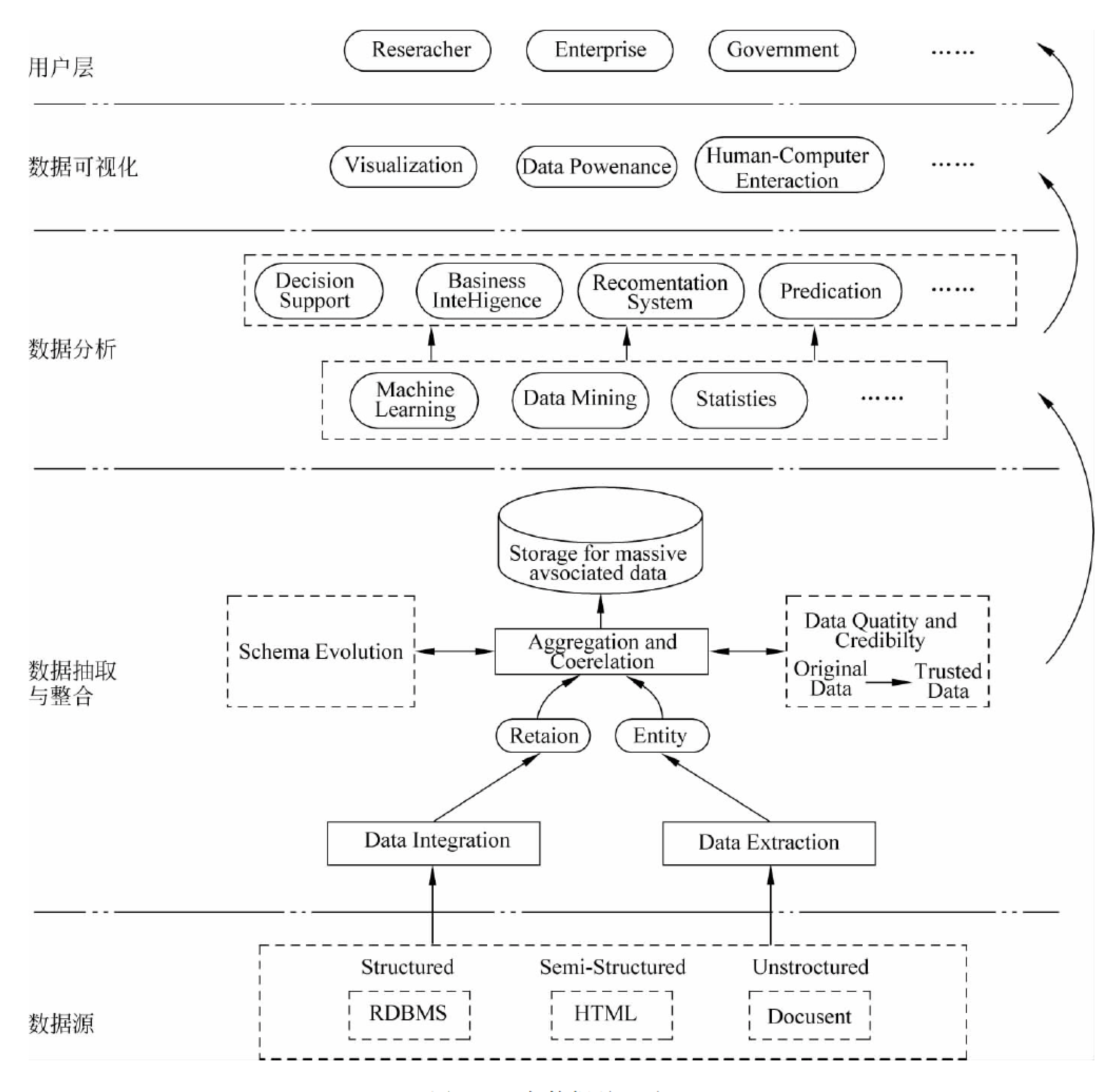


图 4-1　大数据处理流程

可视化：将数据分析结果用直观恰当的方式展现给用户。

应用：面向不同的用户形成应用交互。

我们看到，数据整合是大数据处理流程中非常重要的一环，是数据分析的前提和基础。在数据整合的过程管理中，需要注重以下几个方面。

1．统一模式，制定标准

大数据时代，政府和企业进行数据整合的关键是要形成对数据资源的统一管理和标准建设，第 5 章将具体阐述大数据的管理方法和实践。我们需要把各个行业和业务系统中最核心、最基础、最重要的数据（也称主数据），集中进行数据的 ETL（抽取、清洗、转换），制定好数据存储和交换的模式、接口和访问方法，做好元数据管理，严格把握数据质量和标准，能够把统一的、完整的、准确的、具有权威性的主数据提供给上层需要使用这些数据的模块和应用。

2．构建适合结构化和非结构化数据融合的数据模型

在大数据时代，政府和企业的数据资产不仅局限于原来的结构化文本数据，各种数字化

的音频、视频、图片、邮件、社会化网络、传感器信息等非结构化数据在企业数据资源中的比重逐步攀升。传统的结构化的数据模型和管理方式无法实现对非结构化数据的组织和管理。因此,需要推进结构化和非结构化数据的融合式发展,构建适合结构化和非结构化数据的统一组织和管理的数据模型。实现对海量复杂数据信息的科学有效管理,才能充分挖掘企业数据资源的潜在价值。

3. 注重数据质量和数据可信度

注重数据质量和数据本身的完整性、准确性,才能保障大数据分析结果的真实有效,否则很可能得到错误的分析结果,做出错误的判断和决策,这对于企业的发展是致命的。数据质量和数据可信度就是大数据时代政府和企业的生命线。

4. 重视数据安全管理,确保大数据生态圈信息安全

大数据时代,信息系统之间互联是必然的,它们会形成一个息息相关的生态圈。在这一生态圈里,存储和管理的大量数据信息是企业市场竞争力的核心,需要对数据安全问题进行控制和管理。因此,企业在数据整合过程中应以数据安全管理为前提,需要与上下游企业以及安全管理机构、评测机构等第三方机构开展广泛合作,从企业管理制度、流程和技术手段等多方面协作确保大数据生态圈的数据信息安全。

大数据的整合,除了数据层面的整合,还涉及和数据处理架构的整合,包括和大数据平台以及平台基础设施的整合。大数据平台整合,指的是大数据如何结合大数据处理平台,能够完成数据的统一存储和深度分析、展示及应用;而基础设施的整合,则包含与存储架构的整合、与网络架构的整合,还有与虚拟化技术的整合。下面章节中将具体介绍整合的机制和方式。

4.1　大数据平台整合

第 3 章中介绍了大数据平台的架构体系。大数据平台的整合,涵盖了数据在平台上如何融合基础设施,以及如何和相关的采集、存储、分析、展示交互及应用模块进行接口。下面首先讨论一下最主流的开源大数据平台 Hadoop 的平台整合。

Hadoop 大数据处理平台堪称大数据领域的开山鼻祖,它是 Google 的 GFS 文件系统和 MapReduce 分布式处理框架的开源实现。虽然在此之前有很多类似的分布式存储和计算平台,但真正能实现工业级应用、降低使用门槛、带动业界大规模部署的就是 Hadoop。得益于 MapReduce 框架的易用性和容错性,以及同时包含存储系统和计算系统,使得 Hadoop 成为大数据处理平台的基石之一。Hadoop 能够满足大部分的离线存储和离线计算需求,且性能表现不俗;小部分离线存储和计算需求,在对性能要求不高的情况下,也可以使用 Hadoop 实现。随着整个 Hadoop 开源体系的不断完善和进步,逐步形成了基于 Hadoop 的一个大数据产业生态链。我们以 Cloudera 的 Hadoop 产品链为例说明它的生态构成,如图 4-2 所示。

整个 Hadoop 生态链有以下几个主要组成部分。

(1) Hadoop HDFS 分布式文件系统:能够在大量的存储节点上保存海量数据,并且具备自动备份和容错机制。

(2) MapReduce 分布式计算框架:基于大规模的存储和计算节点进行分布式的数据

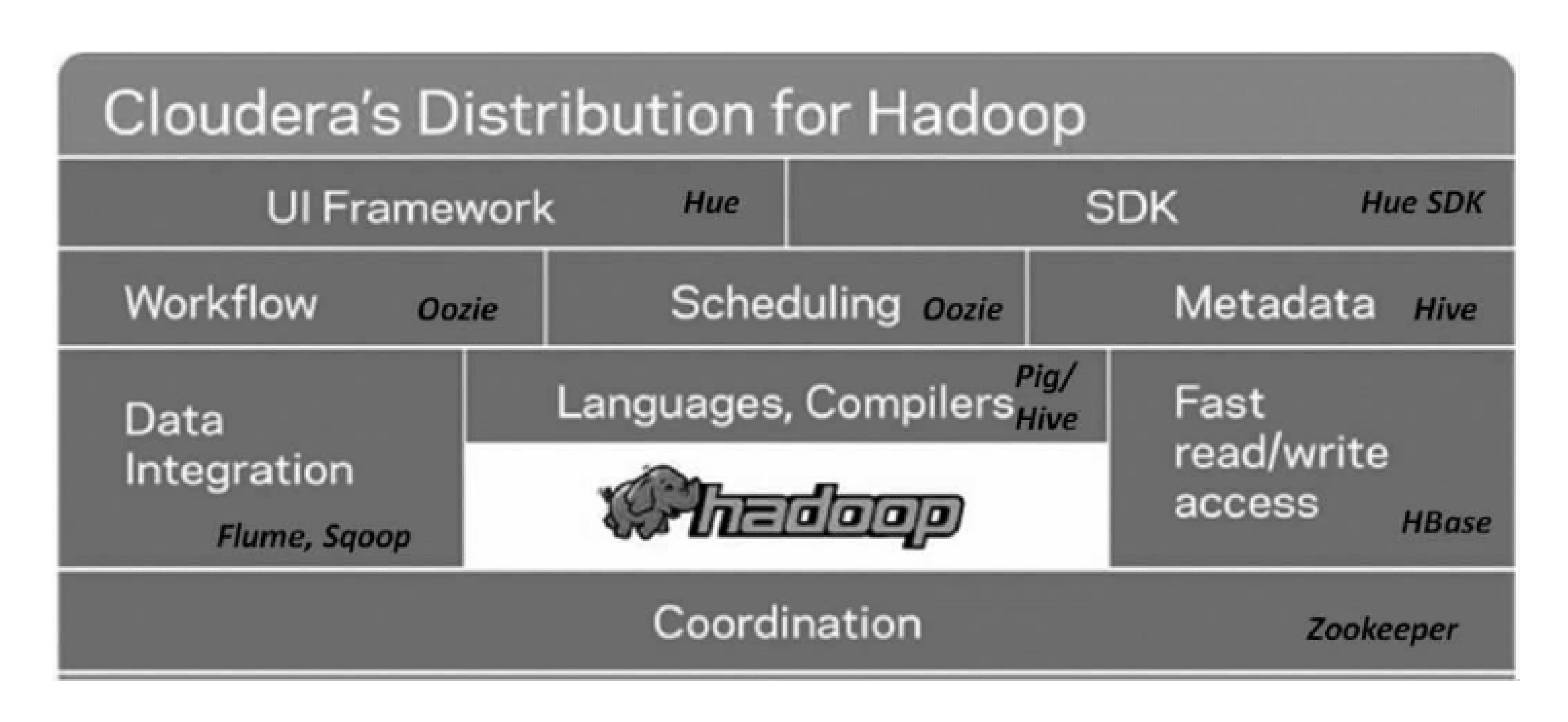


图 4-2 Cloudera 的 Hadoop 大数据生态链

处理。

（3）Zookeeper：集群锁及同步管理工具。

（4）HBase 分布式数据库：能够进行海量数据的数据库管理。

（5）Flume：Sqoop 数据整合/集成工具。

（6）Pig/Hive：基于 Hadoop 的高级分析语言。

（7）Oozie：工作流管理和调度。

（8）Hue：基于交互式图形界面的管理工具。

下面介绍其中几个核心部件。

4.1.1 HDFS 分布式文件系统

Hadoop 由许多元素构成，其最底部是 Hadoop Distributed File System（HDFS），它存储 Hadoop 集群中所有存储节点上的文件，是 GFS 的开源实现，因此本节只对 HDFS 做简要的介绍，包括 HDFS 的特点、架构、读写操作过程。

HDFS 是一种分布式文件系统，运行于大型商用机集群，为 HBase 提供了高可靠性的底层存储支持。由于 HDFS 具有高容错性的特点，所以可以设计部署在低廉的硬件上。它可以以很高的高吞吐率来访问应用程序的数据，适合那些有着超大数据集的应用程序。HDFS 与其他分布式文件系统有许多相似点，但也有几个不同点。一个明显的区别是 HDFS 的"一次写入，多次读取（write-once-read-many）"模型，该模型降低了并发性控制要求，简化了数据聚合性，支持高吞吐量访问。

HDFS 的另一个独特的特性是下面这个观点：将处理逻辑放置到数据附近通常比将数据移向应用程序空间更好（移动程序比移动数据更划算）。通常一个数据处理程序只有几 KB～几 MB 的大小，而数据则非常大，显然，将程序移动到数据所在的位置，处理完数据之后将处理结果传回调用方，这样能节省很多网络带宽资源。

HDFS 将数据写入严格限制为一次一个写入程序。字节总是被附加到一个流的末尾，字节流总是以写入顺序存储。

HDFS 有许多目标，下面是一些最明显的目标。

（1）通过检测故障和应用快速、自动的恢复实现容错性。由于 HDFS 建立在大量普通的硬件设备上，因此硬件故障是常见的问题，整个 HDFS 由数百台或数千台存储着数据文件的服务器组成，而如此多的服务器意味着高故障率，所以故障的检测和自动快速恢复是

HDFS 的一个核心目标。

(2) 通过 MapReduce 流进行数据访问。HDFS 使应用程序能流式地访问它们的数据集。HDFS 被设计成适合进行批量处理,而不是用户交互式的处理。所以它重视数据吞吐量,而不是数据访问的反应速度。

(3) 简单可靠的聚合模型。

(4) 处理逻辑接近数据,而不是数据接近处理逻辑。

(5) 跨异构普通硬件和操作系统的可移植性。

(6) 可靠存储和处理大量数据的可伸缩性。

(7) 通过跨多个普通个人计算机集群分布数据和处理来节约成本。

(8) 通过分布数据和逻辑到数据所在的多个节点上进行平行处理来提高效率。

(9) 通过自动维护多个数据副本和在故障发生时自动重新部署处理逻辑来实现可靠性。

HDFS 是分布式计算的存储基石,Hadoop 的分布式文件系统和其他分布式文件系统有很多类似的特质。分布式文件系统具有以下几个特点。

(1) 对于整个集群有单一的命名空间。

(2) 数据一致性。适合一次写入多次读取的模型,客户端在文件没有被成功创建之前无法看到文件存在。

(3) 文件会被分割成多个文件块,每个文件块被分配存储到数据节点上,而且根据配置会由复制文件块来保证数据的安全性。

4.1.2　MapReduce 分布式计算框架

MapReduce 是一种用于大规模数据集(大于 1TB)的并行运算的编程模型。概念“Map(映射)”和“Reduce(归约)”以及它们的主要思想,都是从函数式编程语言里借用而来的,同时也包含从矢量编程语言里借来的特性。MapReduce 极大地方便了编程人员在不会分布式并行编程的情况下,将自己的程序运行在分布式系统上。

许多人认为这种编程方式的重大变化将带来一次软件的并发危机,因为传统的软件方式基本上是单指令单数据流的顺序执行,这种顺序执行十分符合人类的思考习惯,却与并发并行编程格格不入。基于集群的分布式并行编程,能够让软件与数据同时运行在连成一个网络的许多台计算机上,这里的每一台计算机均可以是一台普通的 PC。这样的分布式并行环境的最大优点是,可以很容易地通过增加计算机来扩充新的计算节点,并由此获得不可思议的海量计算能力,同时又具有相当强的容错能力,一批计算节点失效也不会影响计算的正常进行以及结果的正确性。Google 就是这么做的,他们使用了叫做 MapReduce 的并行编程模型进行分布式并行编程,运行在叫做 GFS(Google File System)的分布式文件系统上,为全球亿万用户提供搜索服务。

Hadoop 实现了 Google 的 MapReduce 编程模型,提供了简单易用的编程接口,也提供了它自己的分布式文件系统 HDFS,与 Google 不同的是,Hadoop 是开源的,任何人都可以使用这个框架来进行并行编程。如果说分布式并行编程的难度足以让普通程序员望而生畏的话,开源的 Hadoop 的出现,则极大地降低了它的门槛。你会发现,基于 Hadoop 编程非常简单,无须任何并行开发经验,也可以轻松地开发出分布式的并行程序,并让其令人难以

置信地同时运行在数百台机器上，然后在短时间内完成海量数据的计算。你可能会觉得你不可能会拥有数百台机器来运行你的并行程序，而事实上，随着"云计算"的普及，任何人都可以轻松获得这样的海量计算能力。例如，现在 Amazon 公司的云计算平台 Amazon EC2 已经提供了这种按需计算的租用服务。

MapReduce 是 Google 提出的一种并行化编程模型，简而言之，它体现了分而治之的策略，它将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数：Map 和 Reduce。适合用 MapReduce 来处理的数据集（或任务），需要满足一个基本要求：待处理的数据集可以分解成许多小的数据集，而且每一个小数据集都可以完全并行地进行处理。一个 MapReduce 作业（job）通常会把输入的数据集切分为若干独立的数据块，由 map 任务（task）以完全并行的方式处理它们。框架会对 map 的输出先进行排序，然后把结果输入给 reduce 任务。通常作业的输入和输出都会被存储在文件系统中。整个框架负责任务的调度和监控，以及重新执行已经失败的任务。通常，MapReduce 框架和分布式文件系统是运行在一组相同的节点上的，也就是说，计算节点和存储节点通常在一起。这种配置允许框架在那些已经存好数据的节点上高效地调度任务，这可以使整个集群的网络带宽被非常高效地利用。MapReduce 框架由单独一个 master JobTracker 和每个集群节点一个 slave TaskTracker 共同组成。这个 master 负责调度构成一个作业的所有任务，这些任务分布在不同的 slave 上，master 监控它们的执行，重新执行已经失败的任务。而 slave 仅负责执行由 master 指派的任务。应用程序至少应该指明输入/输出的位置（路径），并通过实现合适的接口或抽象类提供 map 和 reduce 函数，再加上其他作业的参数，就构成了作业配置。然后，Hadoop 的 job client 提交作业（jar 包/可执行程序等）和配置信息给 JobTracker，后者负责分发这些软件和配置信息给 slave、调度任务且监控它们的执行，同时提供状态和诊断信息给 job client。虽然 Hadoop 框架是用 Java 实现的，但 MapReduce 应用程序则不一定要用 Java 来写。

如图 4-3 所示，下面介绍一下 Hadoop MapReduce 框架的执行原理。

谈 MapReduce 运行机制，可以从很多不同的角度来描述，比如说从 MapReduce 运行流程来讲解，也可以从计算模型的逻辑流程来进行讲解，也许有些深入理解了 MapReduce 运行机制还会从更好的角度来描述，但是讲 MapReduce 运行机制有些东西是避免不了的，就是一个个参入的实例对象，一个就是计算模型的逻辑定义阶段，这里的讲解不从什么流程出发，就从这些一个个牵涉的对象出发，不管是物理实体还是逻辑实体。

首先讲讲物理实体，MapReduce 作业的执行涉及如下 4 个独立的实体。

（1）客户端（Client）：编写 MapReduce 程序，配置作业，提交作业，这就是程序员完成的工作。

（2）JobTracker：初始化作业，分配作业，与 TaskTracker 通信，协调整个作业的执行。

（3）TaskTracker：保持与 JobTracker 的通信，在分配的数据片段上执行 Map 或 Reduce 任务，TaskTracker 和 JobTracker 的不同有个很重要的方面，就是在执行任务时 TaskTracker 可以有很多个，而 JobTracker 则只会有一个（JobTracker 只能有一个，就和 HDFS 里 NameNode 一样存在单点故障，后面讲 MapReduce 2.0 的时候会具体介绍）。

（4）HDFS：保存作业的数据、配置信息等，最后的结果也是保存在 HDFS 上面。

下面从逻辑实体的角度讲解 MapReduce 运行机制，这些按照时间顺序包括：输入分片

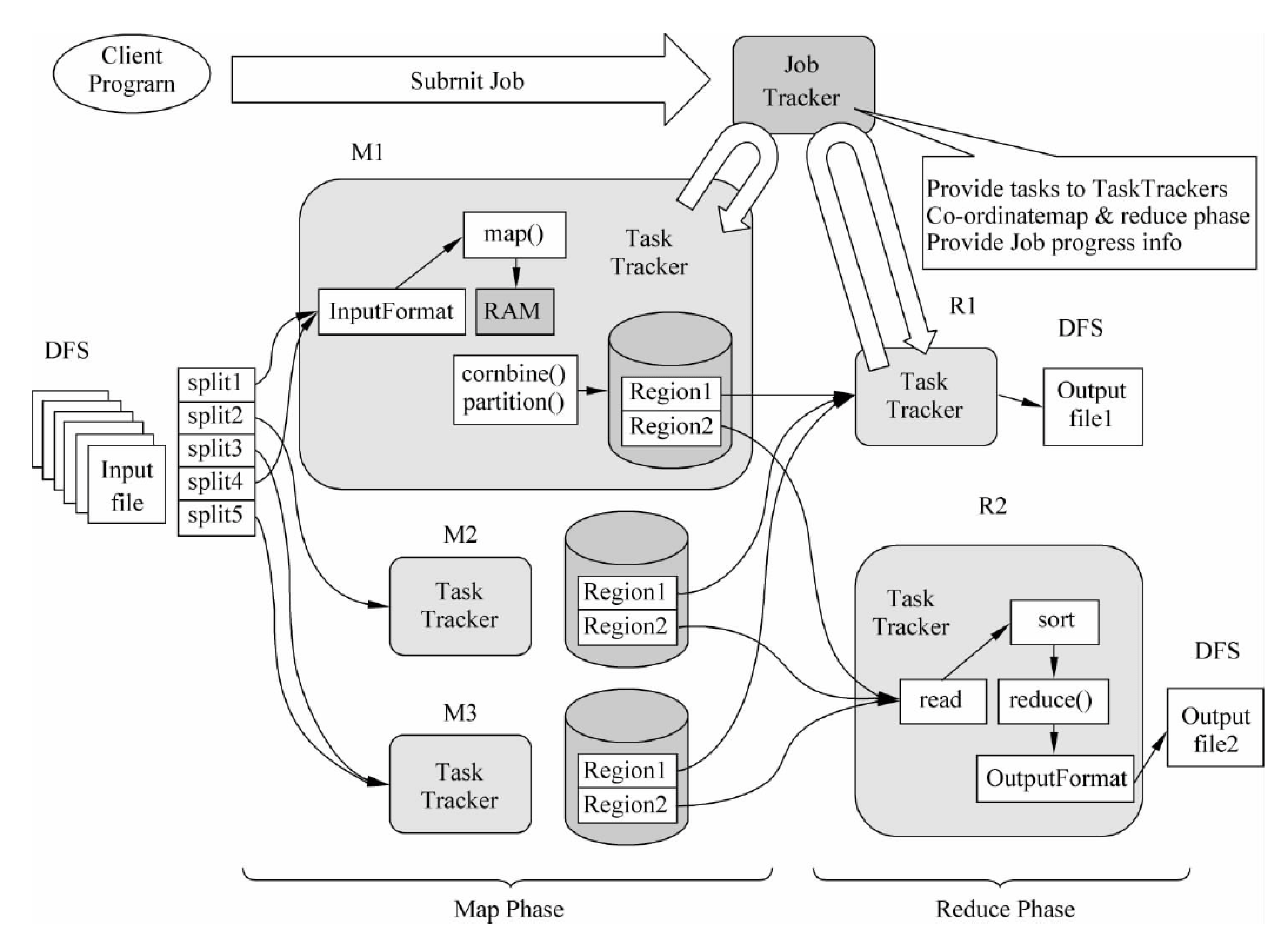


图 4-3 MapReduce 运行原理

(input split)、map 阶段、combiner 阶段、shuffle 阶段和 reduce 阶段。

(1) 输入分片(input split)：在进行 map 计算之前，MapReduce 会根据输入文件计算输入分片，每个输入分片针对一个 map 任务，输入分片存储的并非数据本身，而是一个分片长度和一个记录数据的位置的数组，输入分片往往和 HDFS 的 block(块)关系很密切，假如设定 HDFS 的块的大小是 64MB，如果输入三个文件，大小分别是 3MB、65MB 和 127MB，那么 MapReduce 会把 3MB 文件分为一个输入分片，65MB 则是两个输入分片而 127MB 也是两个输入分片，换句话说，如果在 map 计算前做输入分片调整，例如合并小文件，那么就会有 5 个 map 任务将执行，而且每个 map 执行的数据大小不均，这也是 MapReduce 优化计算的一个关键点。

(2) map 阶段：就是程序员编写好的 map 函数了，因此 map 函数效率相对好控制，而且一般 map 操作都是本地化操作也就是在数据存储节点上进行。

(3) combiner 阶段：combiner 阶段是程序员可以选择的，combiner 其实也是一种 reduce 操作。combiner 是一个本地化的 reduce 操作，它是 map 运算的后续操作，主要是在 map 计算出中间文件前做一个简单的合并重复 key 值的操作，例如对文件里的单词频率做统计，map 计算时如果碰到一个 hadoop 的单词就会记录为 1，但是这篇文章里 hadoop 可能会出现 n 次，那么 map 输出文件冗余就会很多，因此在 reduce 计算前对相同的 key 做一个合并操作，那么文件会变小，这样就提高了宽带的传输效率，毕竟 Hadoop 计算力宽带资源往往是计算的瓶颈也是最为宝贵的资源，但是 combiner 操作是有风险的，使用它的原则是

combiner 的输出不会影响到 reduce 计算的最终输入。例如,如果计算只是求总数、最大值、最小值,可以使用 combiner,但是做平均值计算使用 combiner 的话,最终的 reduce 计算结果就会出错。

(4) shuffle 阶段:将 map 的输出作为 reduce 的输入的过程就是 shuffle 了,这是 MapReduce 优化的重点地方。本节不讲怎么优化 shuffle 阶段,仅讲 shuffle 阶段的原理,因为大部分的书籍里都没讲清楚 shuffle 阶段。shuffle 一开始就是 map 阶段做输出操作,一般 MapReduce 计算的都是海量数据,map 输出时不可能把所有文件都放到内存操作,因此 map 写入磁盘的过程十分复杂,更何况 map 输出时要对结果进行排序,内存开销是很大的。map 在做输出时会在内存里开启一个环形内存缓冲区,这个缓冲区专门用来输出,默认大小是 100MB,并且在配置文件里为这个缓冲区设定了一个阀值,默认是 0.80(这个大小和阀值都是可以在配置文件里进行配置的)。同时 map 还会为输出操作启动一个守护线程,如果缓冲区的内存达到了阀值的 80%时候,这个守护线程就会把内容写到磁盘上,这个过程叫做 spill。另外的 20%内存可以继续写入要写进磁盘的数据,写入磁盘和写入内存操作是互不干扰的,如果缓存区被填满了,那么 map 就会阻塞写入内存的操作,让写入磁盘操作完成后再继续执行写入内存操作。前面讲到写入磁盘前会有个排序操作,这个是在写入磁盘操作时进行,不是在写入内存时进行的。如果定义了 combiner 函数,那么排序前还会执行 combiner 操作。每次 spill 操作也就是写入磁盘操作时就会写一个溢出文件,也就是说在做 map 输出时有几次 spill 就会产生多少个溢出文件,等 map 输出全部做完后,map 会合并这些输出文件。这个过程里还会有一个 partitioner 操作,对于这个操作很多人都很迷糊,其实 partitioner 操作和 map 阶段的输入分片很像,一个 partitioner 对应一个 reduce 作业,如果 MapReduce 操作只有一个 reduce 操作,那么 partitioner 就只有一个,如果有多个 reduce 操作,那么 partitioner 对应地就会有多个,partitioner 因此就是 reduce 的输入分片,这个程序员可以编程控制,主要是根据实际 key 和 value 的值,根据实际业务类型或者为了更好的 reduce 负载均衡要求进行,这是提高 reduce 效率的一个关键所在。到了 reduce 阶段就是合并 map 输出文件了,partitioner 会找到对应的 map 输出文件,然后进行复制操作,复制操作时 reduce 会开启几个复制线程,这些线程默认个数是 5 个,程序员也可以在配置文件中更改复制线程的个数,这个复制过程和 map 写入磁盘过程类似,也有阀值和内存大小,阈值一样可以在配置文件里配置,而内存大小是直接使用 reduce 的 tasktracker 的内存大小,复制时 reduce 还会进行排序操作和合并文件操作,这些操作完了就会进行 reduce 计算了。

(5) reduce 阶段:和 map 函数一样也是程序员编写的,最终结果存储在 HDFS 上。

4.1.3 HBase 分布式数据库

HBase(Hadoop Database)是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统,利用 HBase 技术可在廉价 PC 服务器上搭建起大规模存储集群。HBase 是 Google BigTable 的开源实现,模仿并提供了基于 Google 文件系统的 BigTable 数据库的所有功能。类似 Google BigTable 利用 GFS 作为其文件存储系统,HBase 利用 Hadoop HDFS 作为其文件存储系统;Google 运行 MapReduce 来处理 BigTable 中的海量数据,HBase 同样利用 Hadoop MapReduce 来处理 HBase 中的海量数据;Google BigTable 利用 Chubby 作为协同服务,HBase 利用 Zookeeper 作为协同服务。

在数据存储检索能力方面，HBase 以其优异的随机读写能力著称于世：读方面，根据行键(rowkey)的检索请求响应在 10ms 以内，而根据行键范围的检索请求响应同样在毫秒级；写方面，单条记录的写入响应也在 10ms 左右。

作为一个 NoSQL 数据库，HBase 主要用来存储非结构化和半结构化的松散数据。因此，其本身并未提供 SQL 方面的支持，但其可通过作为其他 SQL 执行引擎(HIVE、SPARK SQL、Phoenix)底层数据源的方式间接满足用户的 SQL 支持需求。

HBase 可以直接使用本地文件系统或者 Hadoop 作为数据存储方式，不过为了提高数据可靠性和系统的健壮性，发挥 HBase 处理大数据量等功能，需要使用 Hadoop 作为文件系统。与 Hadoop 一样，HBase 的目标主要依靠横向扩展，通过不断增加廉价的商用服务器来增加计算和存储能力。HBase 的目标是处理非常庞大的表，可以用普通的计算机处理超过 10 亿行数据并且由数百万列元素组成的数据表。HBase 中的表一般有如下这样的特点。

(1) 大：一个表可以有上亿行，上百万列。

(2) 面向列：面向列(族)的存储和权限控制，列(族)独立检索。

(3) 稀疏：对于为空(null)的列，并不占用存储空间，因此，表可以设计得非常稀疏。

4.1.4　交互式数据查询分析

Hadoop HDFS 中存储了海量数据，可以想象，如果直接访问这些数据将给数据的访问人员带来很大的困难，而且数据的安全性也受到威胁。然而庆幸的是，开源社区专门为 Hadoop 开发了一些交互式数据查询、分析的工具，下面介绍比较常用的。

1. Hive

Hive 是建立在 Hadoop 之上的数据仓库工具，可用于数据集成、ad-hoc 查询、大数据分析。Hive 使用 HDFS 作为数据存储层，提供类似 SQL 的语言(HQL)，将 SQL 语句转换为 MapReduce 任务，通过 Hadoop-MapReduce 完成数据计算；通过 HQL 提供给使用者部分和传统 RDBMS 一样的表格查询特性和分布式存储计算特性。Hive 诞生于 Facebook，Facebook 拥有海量的日志数据，而这里面很大一部分是结构化数据，Hive 以较低的成本完成了以往需要大规模数据库才能完成的任务，并且学习门槛相对较低，应用开发灵活而高效。

2. Pig

Pig 最开始是 2006 年夏天雅虎的一个研究项目，后来发展成为 Hadoop 的一个子项目。它是一个基于 Hadoop 并运用 MapReduce 和 HDFS 实现大规模数据分析的平台，为创建 MapReduce 应用程序提供一种相对简单的工具，为海量数据的并行处理提供了操作及编程的接口。Pig 已经逐渐发展成为能够分析大数据的高级数据流编程语言和执行框架。

Pig 由基础设施和 Pig Latin 编程语言构成。其基础设施可以支持在分布式文件系统上运行应用程序。Pig 使用 Hadoop 框架，因此可以对所有的转换和协调工作进行管理。首先它会自动在 Pig Latin 脚本上执行优化，然后会将相应的操作转换成一个或多个 MapReduce 操作。Pig 会在 Hadoop 集群上运行这些操作，并反映运行状态和错误信息提示等。

4.1.5 数据收集、转换工具

大量数据的收集与转换工作对于 Hadoop 来说也是一件轻松的事，因为它有专门的数据收集、转换工具的支持，大量的数据采集和存储（如日志文件）往往需要经过一系列的处理（数据 ETL），有了这些工具的支持就使得工作得以简化。下面介绍两个常用的工具。

1. Flume

Flume 是一个分布式、可靠、高可用性的海量日志采集、聚合和传输的系统。在其实现架构中，最重要的特点是简单灵活的数据流抽象处理。同时，在 Flume 中，通过 Zookeeper 保证配置数据的一致性和可用性。Flume 具有以下特点。

(1) 可靠性：提供端到端的可靠传输，数据本地化保存等可靠性选项。

(2) 可管理性：通过 Zookeeper 保证配置数据的可用性，并使用多个 master 管理所有节点。

(3) 可扩展性：可以用 Java 语言实现新的自定义功能。

2. Sqoop

Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（MySQL，Oracle，Postgres 等）中的数据导入到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导入到关系型数据库中。具体的整合会在后续章节单独讨论。

4.1.6 其他大数据平台

除了 Hadoop 之外，业界还有实时性更强的大数据处理平台，其中以 Storm 和 Spark 为代表，这里简要介绍一下。

Hadoop 极大降低了海量数据计算能力的门槛，使得各个业务都可以快速使用 Hadoop 进行大数据分析，随着分析计算的不断深入，差异化的需求慢慢浮现了。人们开始发现，某些计算，如果时效性更快，收益会变得更大，能提供给用户更好的体验。一开始，在 Hadoop 平台上为了提高时效性，往往会将一整批计算的海量数据，切割成小时级数据，甚至亚小时级数据，从而变成相对轻量的计算任务，使得在 Hadoop 上可以较快地计算出当前片段的结果，再把当前片段结果跟之前的累积结果进行合并，就可以较快地得出当前所需的整体结果，实现较高的时效性。但随着互联网行业竞争越来越激烈，对时效性越来越看重，尤其是实时分析统计的需求大量涌现，分钟级甚至秒级输出结果，是大家所期望的。Hadoop 计算的时效性所能达到的极限一般为 10min 左右，受限于集群负载和调度策略，要想持续稳定地低于 10min 是非常困难的，除非是专用集群。因此，为了实现更高的时效性，在分钟级、秒级，甚至毫秒级内计算出结果，Storm 应运而生，它完全摆脱了 MapReduce 架构，重新设计了一个适用于流式计算的架构，以数据流为驱动，触发计算，因此每来一条数据，就可以产生一次计算结果，时效性非常高，一般可以达到秒级。而且它的有向无环图计算拓扑的设计，提供了非常灵活丰富的计算方式，覆盖了常见的实时计算需求，因此在业界得到了大量的部署应用。

Storm 的核心框架保证数据流可靠性的方式是：每条数据会被至少发送一次，即正常情况会发送一次，异常情况会重发。这样会导致中间处理逻辑有可能收到两条重复的数据。

大多数业务中这样不会带来额外的问题，或者是能够容忍这样的误差，但对于有严格事务性要求的业务，则会出现问题，例如，扣钱重复扣了两次这是用户不可接受的。为了解决此问题，Storm 引入了事务拓扑，实现了精确处理一次的语义，后来被新的 Trident 机制所取代。Trident 同时还提供了实时数据的 join、group by、filter 等聚合查询操作。

随着大数据平台的逐步普及，人们不再满足于如数据统计、数据关联等简单的挖掘，渐渐开始尝试将机器学习/模式识别的算法用于海量数据的深度挖掘中。因为机器学习/模式识别的算法往往比较复杂，属于计算密集型的算法，且是单机算法，所以在没有 Hadoop 之前，将这些算法用于海量数据上几乎不可行，至少是工业应用上不可行：一是单机计算不了如此大量的数据；二是就算单机能够支撑，但计算时间太长，通常一次计算耗时从几个星期到几个月不等，这对于工业界来说资源和时间的消耗不可接受；三是没有一个很易用的并行计算平台，可以将单机算法快速改成并行算法，导致算法的并行化成本很高。而有了 Hadoop 之后，这些问题迎刃而解，大量机器学习/模式识别的算法得以快速用 MapReduce 框架并行化，被广泛用在搜索、广告、自然语言处理、个性化推荐、安全等业务中。

相比而言，上述的机器学习/模式识别算法往往都是迭代型的计算，一般会迭代几十至几百轮，那么在 Hadoop 上就是连续的几十至几百个串行的任务，前后两个任务之间都要经过大量的 IO 来传递数据。据不完全统计，多数的迭代型算法在 Hadoop 上的耗时，IO 占了 80%左右，如果可以省掉这些 IO 开销，那么对计算速度的提升将是巨大的，因此业界兴起了一股基于内存计算的潮流，而 Spark 则是这方面的佼佼者。它提出了 RDD 的概念，通过对 RDD 的使用将每轮的计算结果分布式地放在内存中，下一轮直接从内存中读取上一轮的数据，节省了大量的 IO 开销。同时它提供了比 Hadoop 的 MapReduce 方式更加丰富的数据操作方式，有些需要分解成几轮的 Hadoop 操作，可在 Spark 里一轮实现。因此对于机器学习/模式识别等迭代型计算，比起 Hadoop 平台，在 Spark 上的计算速度往往会有几倍到几十倍的提升。另一方面，Spark 的设计初衷就是想兼顾 MapReduce 模式和迭代型计算，因此老的 MapReduce 计算也可以迁移至 Spark 平台。由于 Spark 对 Hadoop 计算的兼容，以及对迭代型计算的优异表现，成熟之后的 Spark 平台得到迅速的普及。

人们逐渐发现，Spark 所具有的优点，可以扩展到更多的领域，现在 Spark 已经向通用多功能大数据平台的方向迈进。为了让 Spark 可以用在数据仓库领域，开发者们推出了 Shark，它在 Spark 的框架上提供了类 SQL 查询接口，与 Hive QL 完全兼容，但最近被用户体验更好的 Spark SQL 所取代。Spark SQL 涵盖了 Shark 的所有特性，并能够加速现有 Hive 数据的查询分析，以及支持直接对原生 RDD 对象进行关系查询，显著降低了使用门槛。在实时计算领域，Spark streaming 项目构建了 Spark 上的实时计算框架，它将数据流切分成小的时间片段(例如几秒)，批量执行。得益于 Spark 的内存计算模式和低延时执行引擎，在 Hadoop 上做不到的实时计算，在 Spark 上变得可行。虽然时效性相比专门的实时处理系统有一点儿差距，但也可用于不少实时/准实时场景。另外，Spark 上还有图模型领域的 Bagel，其实就是 Google 的 Pregel 在 Spark 上的实现。它提供基于图的计算模式，后来被新的 Spark 图模型 API——GraphX 所替代。

大数据平台极大地提高了业界的生产力，使得海量数据的整合、存储、计算、分析变得更加容易和高效。通过与这些平台的集成，可以快速实现数据资源的整合，发掘数据的价值。

4.2 大数据与存储架构的整合

4.2.1 传统存储架构

在企业建立初期，用户的数据规模并不大，存储需求也相对简单。人们一般是采用DAS直连存储的架构方案。这种存储方案的服务器结构如同PC架构，外部数据存储设备(如磁盘阵列、磁带机、光盘机等)都直接挂接在服务器内部总线上，数据存储设备是整个服务器结构的一部分。DAS的这种直连方式可以解决单台服务器的存储扩展、高性能传输需求，同时可以构建基于磁盘阵列的双机高可用系统，满足数据高可用的需求。但由于这种存储技术是把设备直接挂在服务器上，随着需求的不断增大，越来越多的设备添加到网络环境中，导致服务器和存储独立数量较多，资源利用率低下，使得数据共享受到严重的限制。因此适用在一些小型网络应用中。

随着企业的发展，应用的复杂度不断加大，需要在不同操作系统间共享文件和应用，并提高性能和存储的扩展性。NAS网络存储技术改进了DAS存储技术，通过标准的拓扑网络，可以无须服务器直接与存储设备连接，不依赖于通用的操作系统，所以存储容量可以很好地扩展，对于原来的服务器性能也没有任何的影响。但是NAS不适合数据库存储(不适合I/O密集型应用)，另外传输速率低成为瓶颈。

NAS存储架构图如图4-4所示。

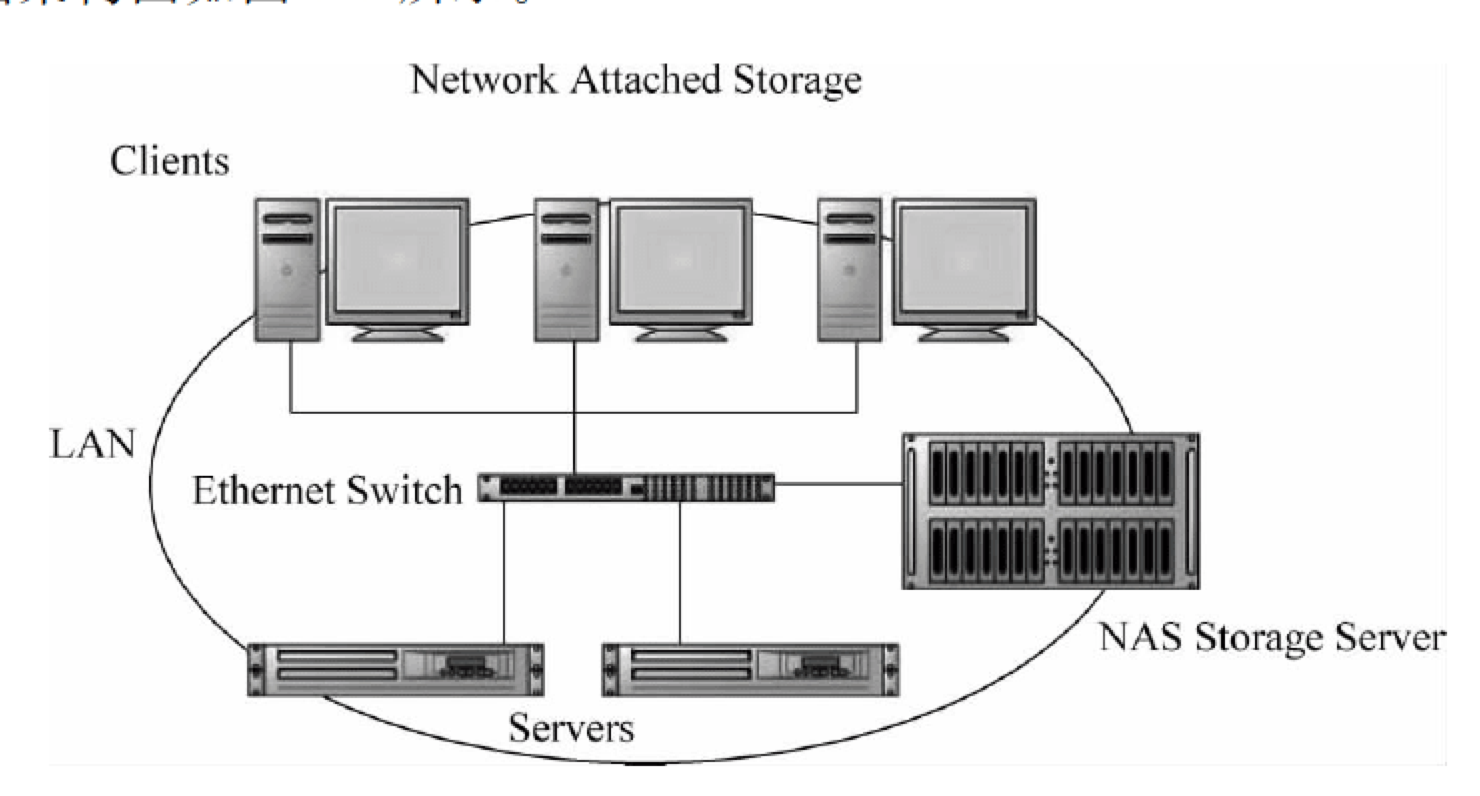


图4-4 NAS网络存储架构图

在企业中，某些核心应用对性能和可靠性有更高的要求，但是NAS存储技术方案的传输速度和效率是有限的。FC光纤通道技术出现后，SAN(存储区域网络)得到了快速发展，在企业中得到了很好的应用。SAN采用高速光纤通道作为传输体，突破传统网络的瓶颈，在服务器与存储设备之间直接高速数据传输，满足了企业对更高性能和可靠性的需求。SAN的架构更适合高端应用领域。SAN的架构图如图4-5所示。

由于大数据技术的发展，传统的存储系统由于没有采用分布式的文件系统，无法将所有访问压力平均分配到多个存储节点，因而在存储系统与计算系统之间存在着明显的传输瓶颈，由此而带来单点故障等多种后续问题，而集群存储正是解决这一问题，满足新时代要求的一剂良药。而传统存储器暴露出的问题也日益明显。

性能问题：由于数据量的激增，数据的索引效率也变得越来越为人们关注。而动辄上

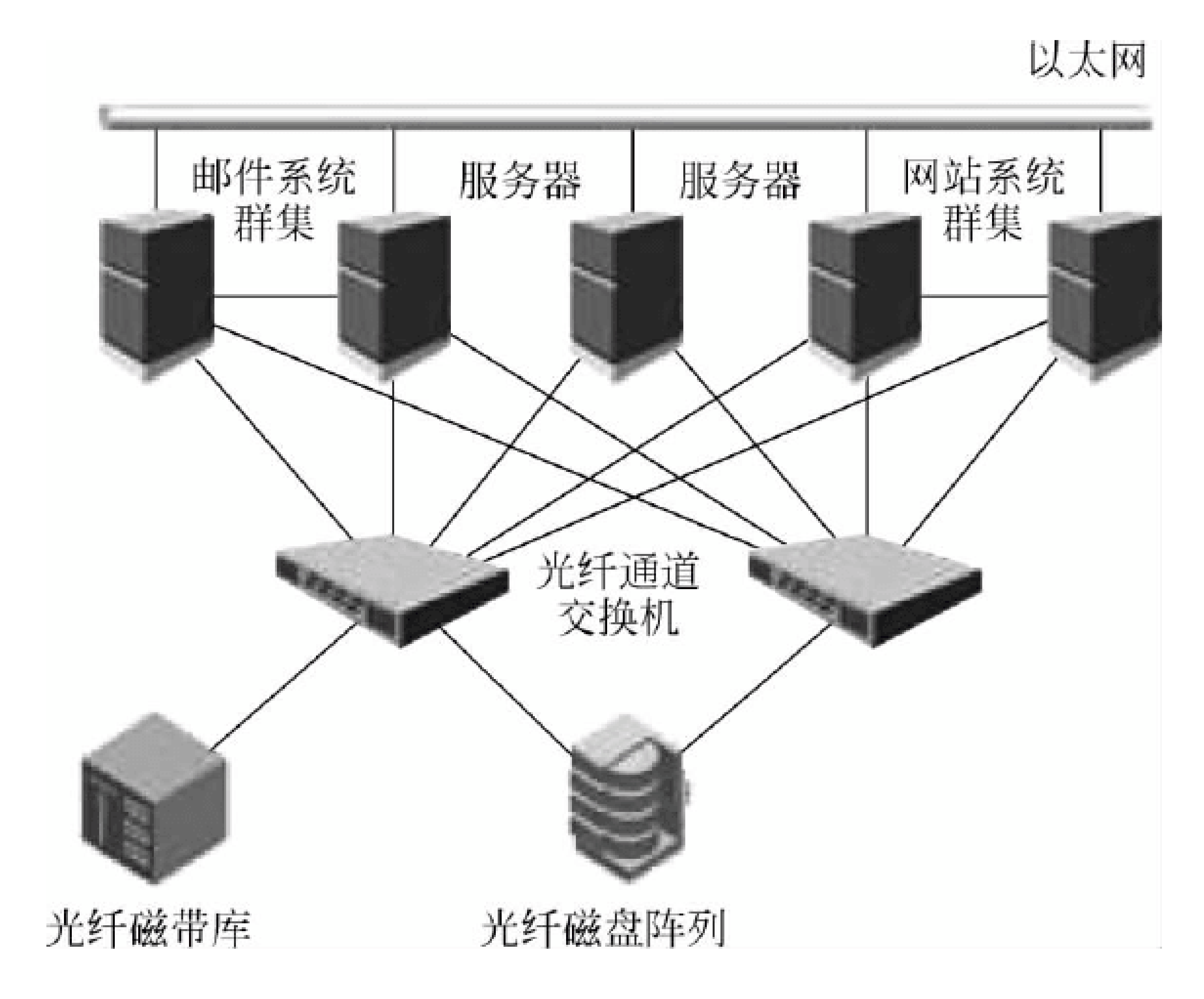


图 4-5　SAN 存储区域网络架构图

TB 的数据，甚至是几百 TB 的数据，在索引时往往需要花上几分钟的时间。

成本激增：在大型项目中，前端图像信息采集点过多，单台服务器承载量有限，就造成需要配置几十台甚至上百台服务器的状况，这就必然导致建设成本、管理成本、维护成本、能耗成本的急剧增加。

磁盘碎片问题：由于视频监控系统往往采用回滚写入方式，这种无序的频繁读写操作，导致了磁盘碎片的大量产生。随着使用时间的增加，将严重地影响整体存储系统的读写性能，甚至导致存储系统被锁定为只读，而无法写入新的视频数据。

4.2.2　集群存储的发展

由于目前一些存储应用受容量可扩展性、性能可扩展性、可用性、可管理性的挑战，“催生”了许多存储集群系统的产生。集群存储是将每个存储设备作为一个存储节点，通过高速互联网络连接起来，将数据分散开存储在多台独立的设备上，这些设备可以独立运作，相互之间又可以合作。每个 I/O 节点不仅可以访问本节点的存储空间，还可以访问其他节点的存储空间。所有存储节点的空间以一个虚拟磁盘的方式提供给客户端用户。组成集群存储可以是块级别的 SAN 集群、文件级别的 NAS 集群和并行文件系统的集群。

集群存储有效地提升了存储设备的容量可扩展性、性能稳定性及系统可管理性。集群存储非常适合那些持续增长的所有规模的不同环境，实现即时供应存储，避免破坏性升级和增加管理的复杂性。在大型数据中心或高性能计算中心的集群存储解决方案，具有高性价比，简单、易于维护，高可靠性/可用性，具有非常高的整合带宽等优点。集群存储最典型的系统是 Google 体系结构，它是大量机器内硬盘的组合，含 899 个机架(每架 80 台 PC，每台 PC 有两个硬盘)，共 79 112 台 PC，有 158 224 个硬盘，总容量为 6180 TB。

近几年逐渐兴起的集群存储技术，不仅轻松突破了 SAN 的性能瓶颈，而且可以实现性能与容量的线性扩展，这对于追求高性能、高可用性的企业用户来说是一个新选择。虽然集群存储在处理非结构化数据方面优势十分明显，但从目前情况来看，集群存储不太可能在短时间内完全取代传统的网络存储方式，SAN 和 NAS 仍会有用武之地。

需要强调的是，虚拟化是实现云计算远景目标的一项核心技术，因为云计算本身就是一个能提供虚拟化和高可用性的新一代计算平台。从目前的市场情况看，服务器虚拟化已经如火如荼，而存储虚拟化的发展相对慢一些。存储虚拟化是一种贯穿于整个 IT 环境、用于简化本来可能会相对复杂的底层基础架构的技术。存储虚拟化的思想是将资源的逻辑映像与物理存储分开，从而为系统和管理员提供一幅简化、无缝的资源虚拟视图。

对于用户来说，虚拟化的存储资源就像是一个巨大的“存储池”，用户不会看到具体的磁盘、磁带，也不必关心自己的数据经过哪一条路径通往哪一个具体的存储设备。这样做的好处是把许多零散的存储资源整合起来，从而提高整体利用率，同时降低系统管理成本。与存储虚拟化配套的资源分配功能具有资源分割和分配能力，可以依据服务水平协议的要求对整合起来的存储池进行划分，以最高的效率、最低的成本来满足各类不同应用在性能和容量等方面的需求。特别是虚拟磁带库，对于提升备份、恢复和归档等应用服务水平起到了非常显著的作用，极大地节省了企业的时间和金钱。

在当今的企业运行环境中，数据的增长速度非常快，而企业管理数据能力的提高速度总是远远落在后面。通过虚拟化，许多既消耗时间又多次重复的工作，例如备份/恢复、数据归档和存储资源分配等，可以通过自动化的方式来进行，大大减少了人工作业。因此，通过将数据管理工作纳入单一的自动化管理体系，存储虚拟化可以显著地缩短数据增长速度与企业数据管理能力之间的差距。

只有网络级的虚拟化，才是真正意义上的存储虚拟化。它能将存储网络上的各种品牌的存储子系统整合成一个或多个可以集中管理的存储池(存储池可跨多个存储子系统)，并在存储池中按需要建立一个或多个不同大小的虚卷，并将这些虚卷按一定的读写授权分配给存储网络上的各种应用服务器。这样就达到了充分利用存储容量、集中管理存储、降低存储成本的目的。

4.2.3 基于 HDFS 的集群存储

前面也多次介绍过，大数据存储系统所基于的是分布式的存储架构。当数据集超过一个单独的物理计算机的存储能力时，便有必要将它分布到多个独立的计算机上。管理着跨计算机网络存储的文件系统称为分布式文件系统。Hadoop 的分布式文件系统称为 HDFS，它是为以流式数据访问模式存储超大文件而设计的文件系统。HDFS 包含几个特点(区别于普通分布式文件系统)：高容错、高吞吐。高容错可以使得系统部署在廉价硬件上，而高吞吐则非常适合做大规模数据集的应用。HDFS 是 Hadoop 应用程序运行的主要分布式存储。一个 HDFS 集群包含一个 NameNode 来管理集群文件系统的元数据，还包含很多 DataNode 来实际存储数据。用户通过在 NameNode 中找到所需访问的文件的元数据，定位到具体存储文件数据块的 DataNode，然后再对数据块进行读取和写入。如图 4-6 所示为 HDFS 的系统结构图。

HDFS 适合做：

(1) 存储大文件，如上 GB、TB 甚至 PB 的大文件。

(2) 一次写入，多次读取。并且每次作业都要读取大部分的数据。

(3) 搭建在普通商业机群上就可以了。虽然这些机器会经常宕机，但 HDFS 有良好的自容错和自恢复机制，不需要人工干预。

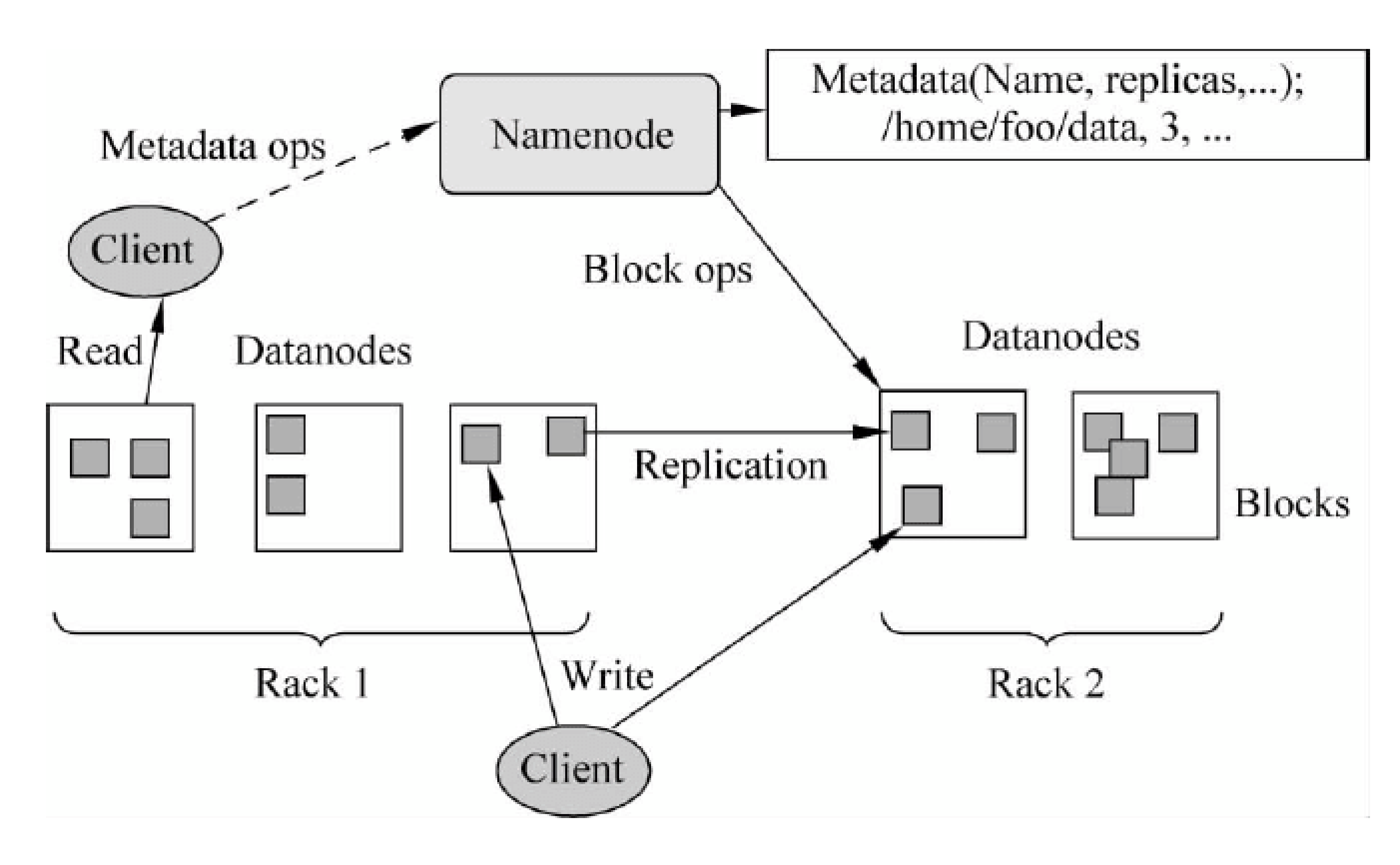


图 4-6　HDFS 系统架构图

HDFS 不适合做：

(1) 实时数据获取。如果有这个需求可以用 HBase 分布式数据库。

(2) 很多小文件的情形。因为 NameNode 要存储 HDFS 的元数据（比如目录的树状结构，每个文件的文件名、ACL、长度、owner、文件内容存放的位置等信息），所以 HDFS 上文件的数目受到 NameNode 内存的限制。

(3) 并发环境下的写入和修改。

因此需要根据具体的应用场景来选择适合该场景的存储架构。总的来说，在如今大数据爆炸的时代，对于如何应对 PB 级别的数据，HDFS 分布式存储是这个时代的产物，是对大数据进行存储处理的优秀架构。

4.2.4　固态硬盘对内存计算的支持

固态硬盘（Solid State Disk 或 Solid State Driver，SSD）是一种以内存作为永久性存储器的计算机存储设备。虽然 SSD 已不是使用“碟盘”来记存数据，而是使用 NAND Flash，但是人们依照命名习惯，仍然称其为固态硬盘或固态驱动器。当然，SSD 内也没有用来驱动旋转的马达。

1. 分类

(1) 易失性内存。由易失性内存制成的固态硬盘主要用于临时性存储。因为这类内存需要靠外界电力维持其记忆，所以由此制成的固态硬盘还需要配合电池才能使用。易失性内存，例如 SDRAM，具有访问速度快的特点。利用这一特点，可以将需要运行的程序从传统硬盘复制到固态硬盘中，然后再交由计算机运行，这样可以避免由于传统硬盘的引导延迟、搜索延迟等对程序以及系统造成的影响。

由易失性内存制成的固态硬盘通常会依靠电池来保证完成应急备份：当电源意外中断时，靠电池驱动的这类固态硬盘可以有足够的时间将数据转移到传统硬盘中。当电力恢复后，再从传统硬盘中恢复数据。

(2) 非易失性内存。非易失性内存的数据访问速度介于易失性内存和传统硬盘之间。和易失性内存相比，非易失性内存一经写入数据，就不需要外界电力来维持其记忆，因此更适于作为传统硬盘的替代品。

闪存当中的NAND Flash是最常见的非易失性内存。小容量的NAND闪存可被制作成带有USB接口的移动存储设备，亦即人们常说的“U盘”。随着生产成本的下降，将多个大容量闪存模块集成在一起，制成以闪存为存储介质的固态硬盘已经是目前的趋势。

目前用来生产固态硬盘的NAND Flash有三种，分别是单层式存储(SLC)、多层式存储(MLC，通常用来指称双层式存储)、三层式存储(TLC)。有些厂商也称TLC为3-bit MLC。SLC、MLC、TLC的读写速度依序从快至慢(约4∶2∶1)，使用寿命依序从长至短(约6∶3∶2)，成本依序从高至低，需要纠错比特数(ECC)则是相反地从低至高(同一制程下1∶2∶4。不过ECC也受制程的影响，同一种芯片，越小尺度的制程需要越多的纠错比特)。固态硬盘的主流从SLC芯片转到MLC芯片，促成了2011年的大降价，固态硬盘因此普及。

由于因为SLC的速度较快但成本过高，用于服务器的企业级SSD都改用了MLC。TLC因为速度较慢但成本低，原本只用来做U盘；不过2012年下半年，SAMSUNG首先推出使用TLC的消费级固态硬盘(型号840系列)，固态硬盘名牌Plextor也打算于2013年量产TLC产品作为低级廉价市场的主力，然而TLC的寿命、速度和可靠性(错误率)成为消费者的最大疑虑。生产商会在TLC SSD使用更先进的主控及更多预留空间(OP)来处理这些问题。

TLC的错误率已经很高，需要使用先进的主控及大量的空间进行纠错。如果发展4-bit MLC会令错误率升得更高，同时寿命更短。三星已量产两代3D垂直闪存，利用3D堆栈增加存储密度。

2. 优点

和传统硬盘相比，固态硬盘具有低功耗、无噪声、抗震动、低热量的特点。这些特点不仅使得数据能更加安全地得到保存，而且也延长靠电池供电的设备的连续运转时间。例如，三星电子于2006年3月推出的容量为32GB的固态硬盘，采用和传统微硬盘相同的1.8英寸规格。其耗电量只有常规硬盘的5%，写入速度是传统硬盘的1.5倍，读取速度是传统硬盘的3倍，并且没有任何噪声。其后固态硬盘取得了飞速的发展。2015年，三星在“闪存高峰会”上发表容量高达16TB的2.5英寸固态硬盘(SSD)PM1633a，其存储容量甚至高过于传统硬盘，接口是SAS 16Gb/s。最初的固态硬盘容量少、价钱高，性价比远不及传统的机器性硬盘。但随着固态硬盘的不断发展，固态硬盘的容量已有实用性，价钱明显下滑之下，已为传统硬盘市场制造危机。

内存作为计算机的重要配件之一，它是硬盘与CPU之间进行沟通的桥梁。它主要用于暂时存放CPU中的运算数据，以及与硬盘等外部存储器交换的数据。由于计算机中所有程序的运行都是在内存中进行的，因此内存的性能以及稳定运行对计算机的影响非常大。在如今数据量如此庞大的情景下，我们在关注大数据集存储的同时还要关注对数据处理的速度，在内存资源有限和成本高昂的情况下，有了SSD对内存计算的支持，我们对大数据进行处理的速度又有了进一步的提升。所以，SSD是在我们做大数据存储时的一种选择。相信在不久的将来，随着SSD技术的进一步发展，对数据的读取写入的速度会得到更大的提高，从而我们在大数据处理上的效率会得到更大的提升。

4.3 大数据与网络架构的发展

大数据技术平台是基于现有网络架构来实现分布式计算的。如图 4-7 所示是一个典型的云计算环境下的企业级网络架构。篇幅所限，这里不展开详细介绍。

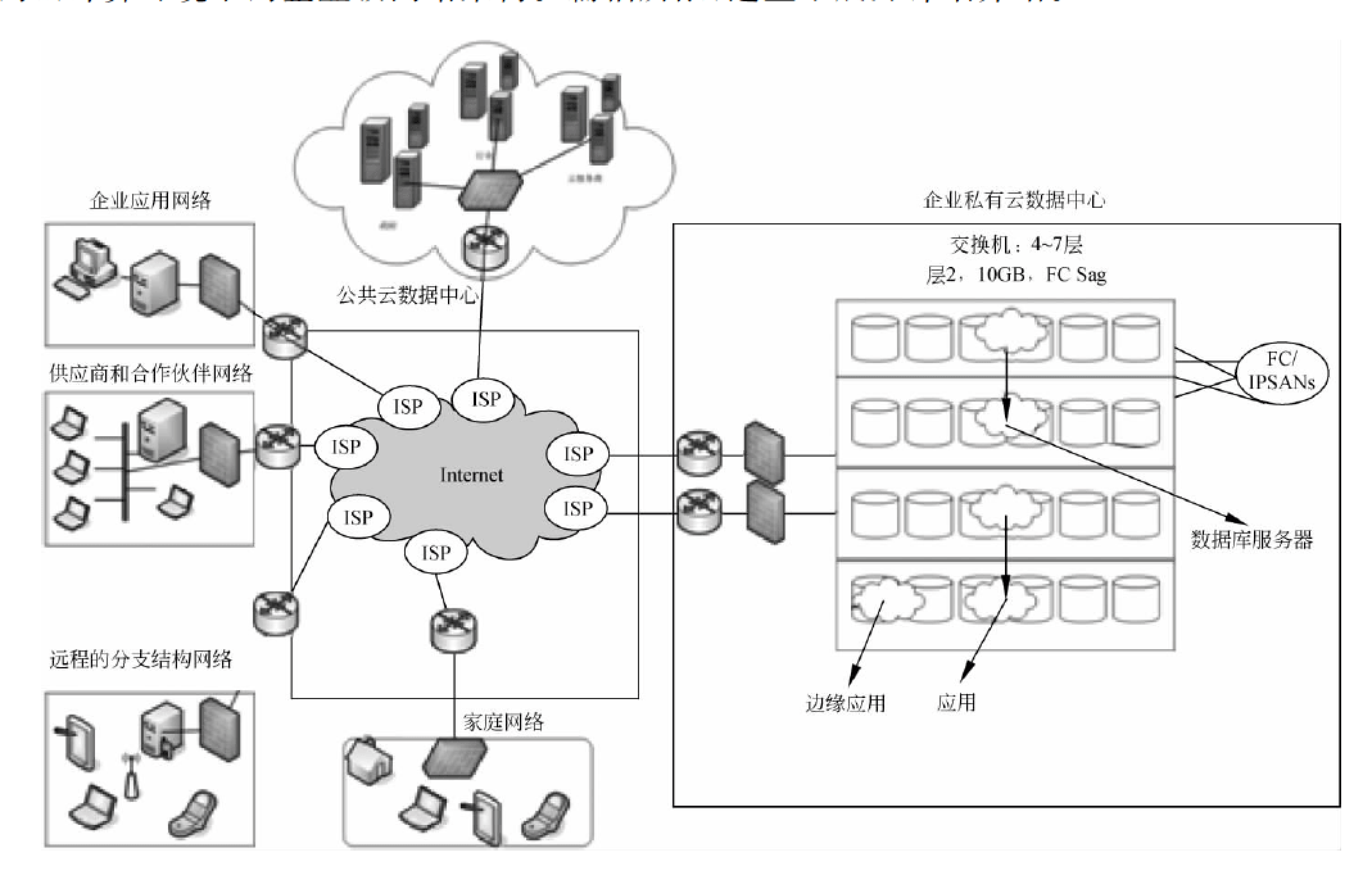


图 4-7 典型的云计算环境下的企业级网络架构

为了进一步满足大数据应用持续的要求，需要对现有企业网络架构进行升级，思科公司提出的统一的以太网结构（Unified Ethernet Fabrics，UEF）或正在兴起的软件定义网络（Software Defined Network，SDN）是解决这个问题的技术趋势。

1. 统一的以太网结构

统一的以太网架构（UEF）正快速发展，它很适合云计算和大数据的需求。UEF 是一个更扁平和集中的网络，它是架构在各种网络设备上的一个虚拟化网络平台。UEF 的特点如下。

（1）集中的网络架构。减少了网络设备的复杂性，以及与多个 Fabrics、分开的网络适配器和布线相联系的大量成本花费。

（2）网络扁平化。网络架构的扁平化设计最大化地提升了网络效率、减少了拥塞，并通过产生用于负载均衡和冗余的第二层网络路径，解决了扫描树的限制。

（3）虚拟化。UEF 通过虚拟底盘的体系架构，统一了多个交换机的访问，逻辑上这些设备被当作一个设备来管理。这就产生了虚拟交换机的资源池，免除了手动配置的必要。这个设计提供了任何设备延迟的可预测的大数据集群服务器之间的流量带宽。

（4）多个路由路径选择。通过利用通过网络的多个路径并连续决定最有效的路由，UEF 能实现全链接的利用。

（5）可靠性。UEF 带来了分布式网络，对失效更有弹性和容错能力。

2. 软件定义网络

软件定义网络(SDN)能够将网络控制从物理基础设施中解耦出来,通过软件和虚拟化在更加全局的角度对网络设备进行控制。不同的网络设备通过开放的接口来进行整合,如OpenFlow,一个可扩展的、可能是开源的网络操作系统架构在 OpenFlow 交换机上,通过很好定义的 API 实现网络操作系统对应用的支撑。SDN 和 OpenFlow 标准被认为是网络领域中的重要发展趋势,它们已经成为谷歌、Facebook 和雅虎等云服务提供商和大型网络公司简化或自动化网络配置的一种主流趋势。用户不再需要手动操作网络中的任何交换机或路由器,即可快速添加和配置更多的网络功能。

SDN 是一个新的网络结构,通过将传统网络设备紧耦合的架构分解为应用、控制、转发独立的三层,并实现可编程控制。传统网络每个路由或者交换设备都是一个控制和数据的合体,数据包由分布式设备自行决定操作方式,最后通过各个设备的合作达到目的。在SDN 架构中,控制层和数据层分离,数据层设备只管数据的转发操作,控制功能被集中转移到称为控制器的服务器,高层的应用、底层的转发设备被抽象为多个逻辑实体。控制器负责收集全网信息,进行决策,并向数据层设备下发策略,每个数据层设备按照这些策略对不同数据包执行操作,未来网络架构如图 4-8 所示。

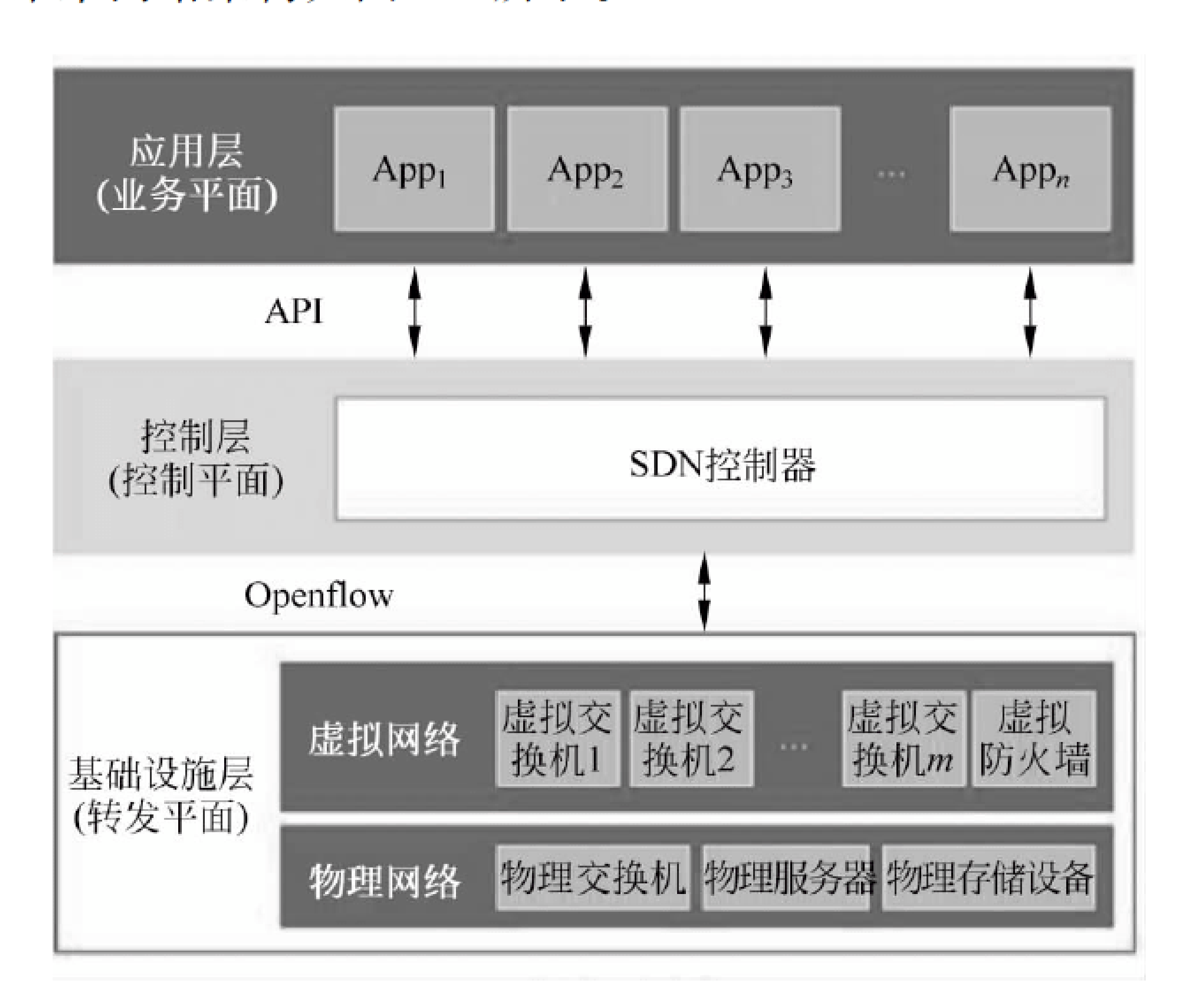


图 4-8 未来网络架构

为了便于第三方通过编程方式对网络资源进行动态分配,需要对网络进行抽象以屏蔽底层复杂度,为上层提供简单的、标准化的、高效的网络资源逻辑或者虚拟实体：第一,使第三方编程可以独立于复杂的物理网络结构；第二,提高网络资源在整网角度的利用率；第三,实现网络资源的快速和动态部署,提高网络的弹性,降低网络调度的颗粒度。因此,在未来网络架构中,采用类似于云架构将计算和存储资源虚拟化的方式,应用 NFV 技术实现网络资源的虚拟化。

由此可见,未来网络架构增强了网络的可控性,降低了网络资源管理的颗粒度,增强了网络的弹性,提高了网络资源的利用率,第三方可以根据不同业务需求对网络资源进行动态和实时的调配,网络资源实现了虚拟化和切片化(类似于虚拟专用网),总的来说,这种架构

让客户拥有更多的控制能力，继承了互联网开放创新的基因，可望突破互联网产业发展瓶颈。

4.4　大数据与虚拟化技术的整合

在计算机中，虚拟化（Virtualization）是一种资源管理技术，是将计算机的各种实体资源，如服务器、网络、内存及存储等，予以抽象、转换后呈现出来，打破实体结构间的不可切割的障碍，使用户可以用比原本的组态更好的方式来应用这些资源。这些资源的新虚拟部分是不受现有资源的架设方式，地域或物理组态所限制的。一般所指的虚拟化资源包括计算能力和资料存储。在实际的生产环境中，虚拟化技术主要用来解决高性能的物理硬件产能过剩和老的旧的硬件产能过低的重组重用，透明化底层物理硬件，从而最大化地利用物理硬件。

虚拟化技术与多任务以及超线程技术是完全不同的。多任务是指在一个操作系统中多个程序同时运行；而在虚拟化技术中，则可以同时运行多个操作系统，而且每一个操作系统中都有多个程序运行，每一个操作系统都运行在一个虚拟的 CPU 或者是虚拟主机上；而超线程技术只是单 CPU 模拟双 CPU 来平衡程序运行性能，这两个模拟出来的 CPU 是不能分离的，只能协同工作。

虚拟化技术是一套解决方案。完整的情况需要 CPU、主板芯片组、BIOS 和软件的支持，例如 VMM 软件或者某些操作系统本身。即使只是 CPU 支持虚拟化技术，在配合 VMM 的软件情况下，也会比完全不支持虚拟化技术的系统有更好的性能。虚拟化技术一般分为全虚拟化和半虚拟化技术。

1. 全虚拟化

全虚拟化（Full Virtualization）也称为原始虚拟化技术，如图 4-9 所示。该模型使用虚拟机协调 guest 操作系统和原始硬件，VMM 在 guest 操作系统和裸硬件之间用于工作协调，一些受保护指令必须由 Hypervisor（虚拟机管理程序）来捕获处理。

全虚拟化的运行速度要快于硬件模拟，但是性能方面不如裸机，因为 Hypervisor 需要占用一些资源。

2. 半虚拟化

半虚拟化（Para Virtualization）是另一种类似于全虚拟化的技术，如图 4-10 所示。它使用 Hypervisor 分享存取底层的硬件，但是它的 guest 操作系统集成了虚拟化方面的代码。该方法无须重新编译或引起陷阱，因为操作系统自身能够与虚拟进程进行很好的协作。

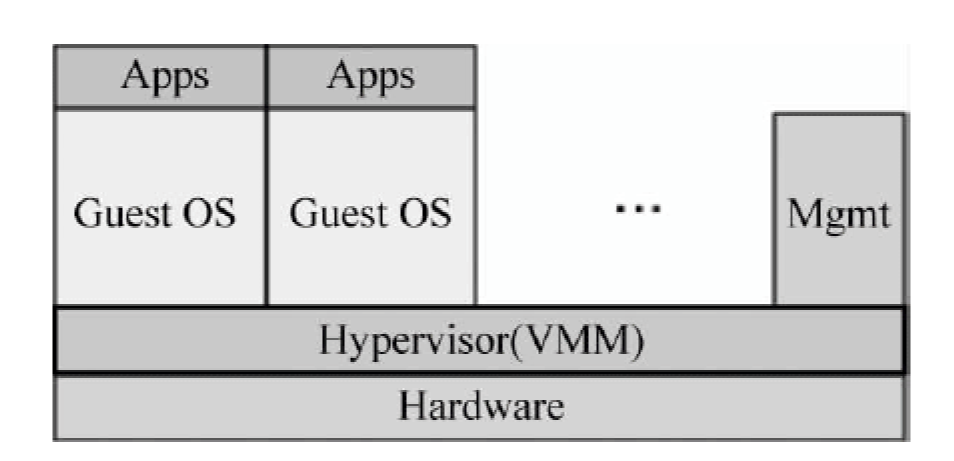


图 4-9　全虚拟化模型

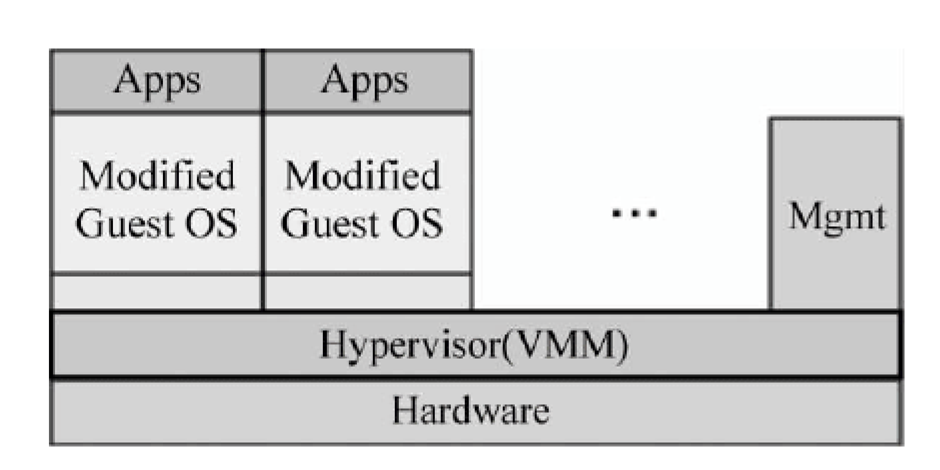


图 4-10　半虚拟化模型

半虚拟化需要 guest 操作系统做一些修改，使 guest 操作系统意识到自己是处于虚拟化环境的，但是半虚拟化提供了与原操作系统相近的性能。

大数据的虚拟化是当前大数据以及 Hadoop 社区的一个发展趋势。随着全球企业 IT 虚拟化的比例突破三分之二，以虚拟化为基础的软件定义的数据中心对企业来讲变得越来越普及和重要，大数据在这样的浪潮下如何影响和融入现有企业数据中心的基础架构变成了现实的挑战。

(1) 虚拟化能够显著提升服务器的利用率，通过整合服务器资源达到更佳的利用率。

(2) 以 x86 服务器为代表的虚拟化本身的拥有成本相对小型计算机和软硬件一体设备来讲，更经济，而且性能表现一点儿也不逊色，横向扩展更是巨大优势。

(3) 虚拟化在云计算(无论是公有云还是私有云)中承担着很重要的基础工作。没有虚拟化技术，云计算的弹性和多租户往往难以得到真正落实。

(4) 虚拟化已经可以支撑企业关键应用(如 ERP、邮件服务器、业务生产数据库等)，这证明在虚拟化和性能稳定性之间已经不再需要二选一。虚拟化迈向全面成熟的标志已经树立。

显然企业虚拟化的进程不会停止，目前包括 VMware 在内的领导厂商都在拓展虚拟化 2.0。不仅是服务器(计算资源)虚拟化，包括存储和网络等过去相对难以直接被虚拟化所用的孤岛都出现了最前沿的创新推动，例如“软件定义数据中心”“存储虚拟化”“网络虚拟化”等热点，都已经出现了具体的产品和解决方案。

大数据的虚拟化，是将大数据的工作负载运行或迁移到虚拟化的基础环境中。除了自然地继承以上所谈到的虚拟化的普遍优点，值得一提的有以下几个特殊的好处。

(1) 由于大数据基础架构在起步时往往难以确定需要多少计算和数据节点，这些节点用物理服务器需要一一去堆。如果没有专家团队支持，将会非常耗时费力，而且将来扩展非常不方便，利用率极低，管理效率问题相当突出。虚拟化不仅可以快速部署集群，更可以灵活管理它们，同时显著提高利用率。

(2) 大数据混合使用共享存储和本地存储，用来提高性能。虚拟化可以完全满足这些需求，并且让我们灵活地扩展和设计策略。

(3) 虚拟化可以将大数据从底层向上对外形成多租户和数据分析服务，很好地隔离计算环境，为推动大数据即服务奠定基础。

(4) 虚拟化还有利于整合和集成其他的数据应用在统一的虚拟化平台上，大大降低 IT 基础架构的复杂度和运维成本。

存储虚拟化，可以形成统一的存储池，屏蔽各个存储设备的异构，实现阵列高可用，以及在线数据迁移等。对于大量的非结构化的数据的存储，通过存储的虚拟化网关，用户不再关心文件存储的路径，通过单一位置提供的文件名就可以访问。存储虚拟化是构建集群存储的基础，能够支持实现海量数据的动态分级存储。

网络虚拟化，可以将两台或多台设备虚拟为一台设备，实现统一转发、统一管理，并实现跨设备的链路捆绑，简化网络协议的部署，大大缩短设备和链路收敛时间(毫秒级)，以链路负载分担方式工作，利用率大大提升。网络虚拟化是实现 SDN 的基础技术。

主机的虚拟化技术，可以实现“一分多”，即将一台服务器虚拟成多台虚拟机，在进行 Hadoop 平台的安装和实验阶段，可以采取这样的方法来进行。主机虚拟化技术也可以实

现“多合一”，将多台服务器虚拟成一个虚拟服务器，实现不同大数据计算集群的统一关联和资源共享。如图 4-11 所示为用虚拟化技术统一大数据平台，可以使得计算资源池能够按需求更快、更容易地提供新的数据集群，允许工作负载的混合，利用虚拟机来提供隔离，基于虚拟的拓扑来优化数据性能，基于虚拟拓扑使得系统更可靠。

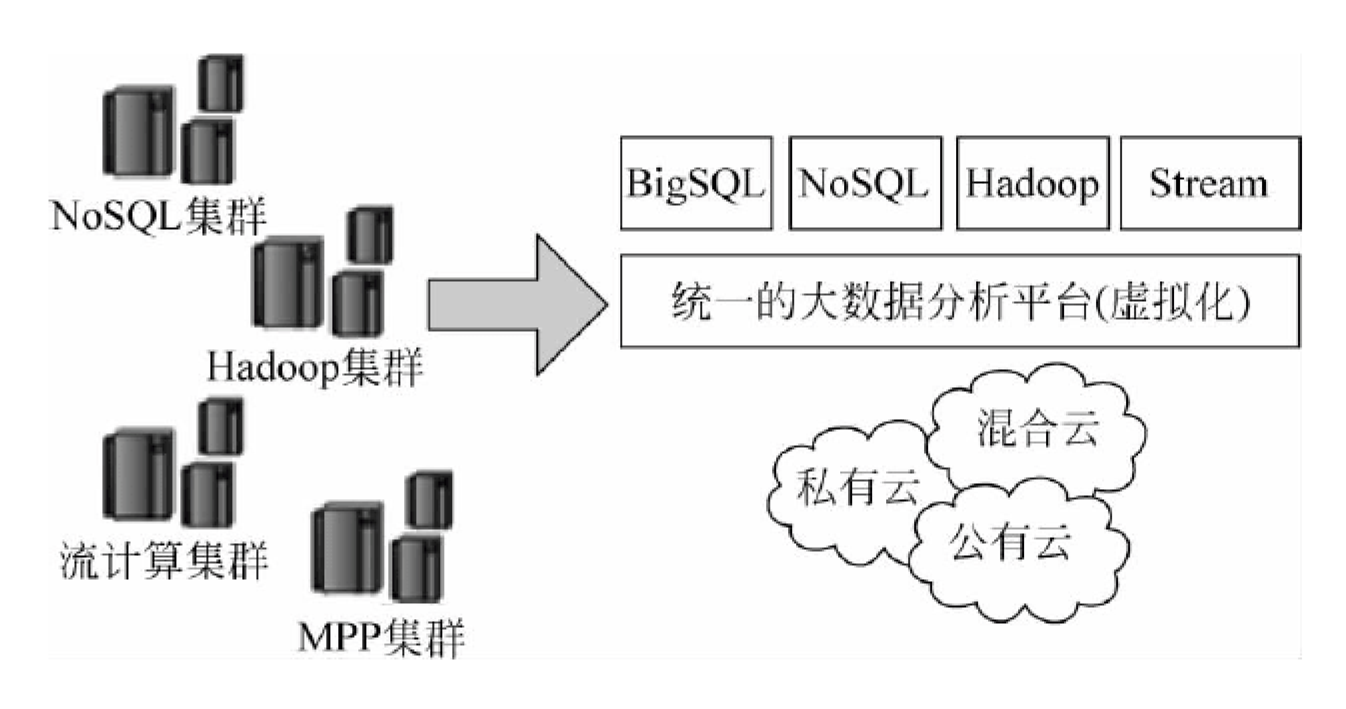


图 4-11　大数据分析平台(虚拟化)

4.5　Hadoop 环境下的数据整合

4.5.1　Hadoop 计算环境下的数据整合问题

随着企业业务的增长，伴随企业各类应用系统逐步启用，结果导致数据量几何级数的增长，传统的整合数据的方式正在受到挑战，与此同时，云计算及网上应用在企业内部产生各类结构化、非结构化数据，这些数据所蕴含的信息(尤其是非结构化数据)是传统分析工具无法捕捉的。

从根本来说，企业信息化的目的是为了降低沟通成本、提高工作效率、增强科学决策能力，从手段上是将分散、无序、无时效的数据变成有序、可分享、有时效、可追溯的数据，前者数据过渡到后者数据，就是无信息(或不可信信息)变成可信信息的过程。数据蕴含的信息有两类：①交易信息，即某一条/或几条数据本身所包含的信息；②统计信息，即数据集合所蕴含的规律性信息。如图 4-12 所示，表现了交易数据与统计数据的关系和传统架构方法，即 ETL 模型。

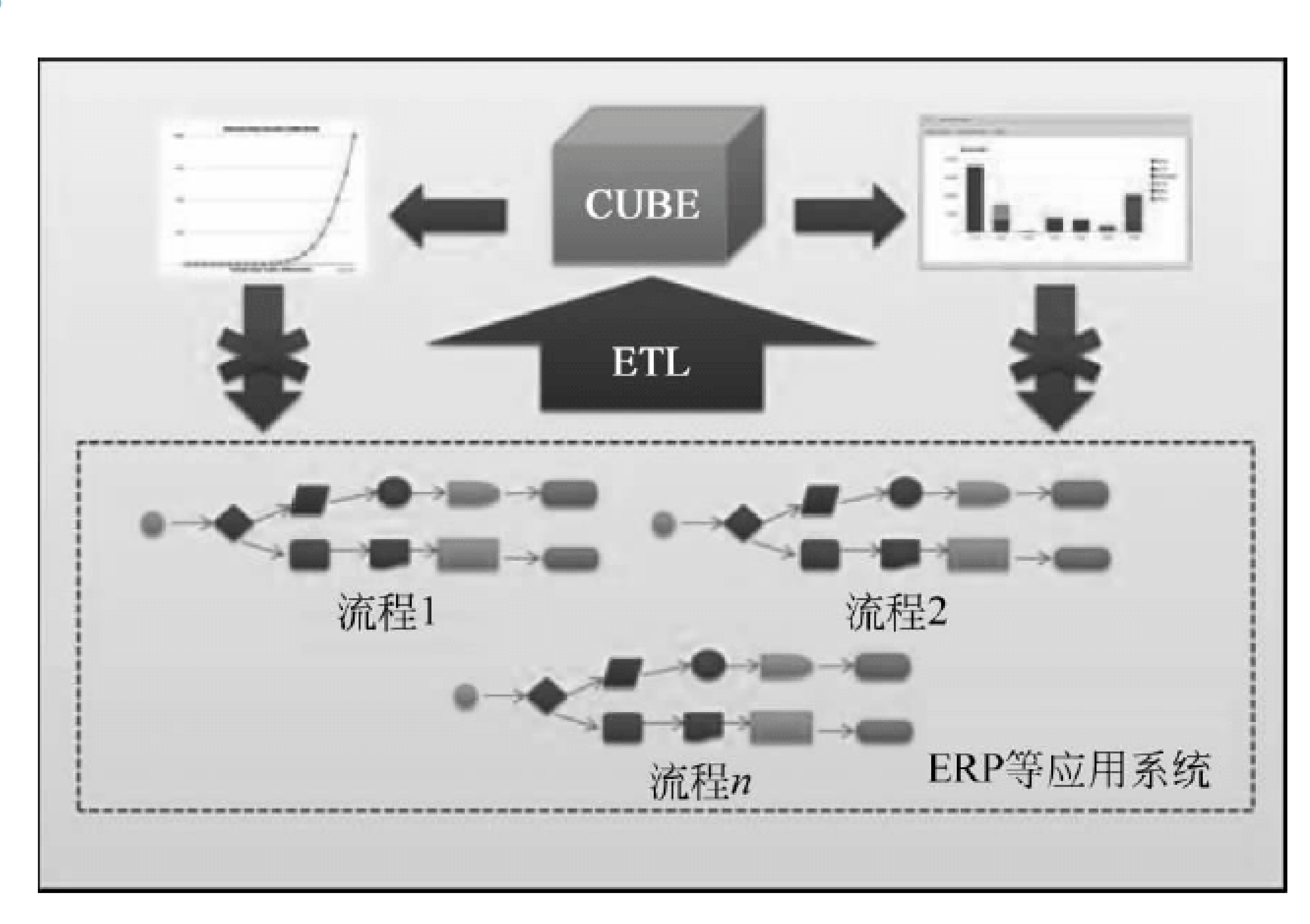


图 4-12　典型传统数据仓库架构

传统整合基本上是基于ETL模式，即从企业内部的信息系统中抽取(Extract)，然后根据预先定义的方式转换(Transform)，最后载入到企业的数据仓库(Load)，大部分企业的ETL程序定义在每天晚上运行，这类方法有以下问题。

(1) 数据仓库的数据不是实时的信息。

(2) 如果内部信息系统数据量很大，ETL处理时间不可能按时完成。

(3) 数据仓库的信息无法快速反馈数据到基层处理商务的人员。

(4) ERP本身在多年数据积累后，事务处理与订单查询都会变慢。

(5) 无法处理大数据，ETL的整个数据处理过程都是建立在已知/预定义的模型之上的，也就是ETL无法发掘到数据集蕴含的未知规律。

(6) 结构化大数据，除上述第(2)点外，针对大数据的深度挖据分析能力(非简单根据预先设计的模型做数据转换)，传统的系统架构中是无法完成的。

(7) 非结构化、半结构化大数据。非架构化数据从本质上来讲，是企业无法预先定义规则的数据类型，据IDC的一项调查报告中指出：企业中80%的数据都是非结构化数据，这些数据每年都按指数增长60%。传统的方式无法计算统计非结构化数据。

Hadoop计算环境下的数据整合问题可以分为两个方面：一方面是整合传统数据源(例如MySQL、Oracle这类的传统关系型数据库)；另一方面则是构建在HDFS之上的数据源间的整合。

我们知道，Hadoop计算环境的底层存储HDFS有别于传统的分布式文件系统。因此，如果要将传统关系型数据库中的数据抽取到HDFS上，需要导出数据文件并调用HDFS提供的API接口实现文件上传(而非传统的复制粘贴文件)。为了简化这个步骤，Apache Sqoop为我们提供了一套简单易用且兼具灵活性的数据整合工具。

那本身就以HDFS为底层存储的数据源的数据整合又是如何解决的呢？考虑到这些数据源的数据文件均存放在HDFS上，如果要整合这些数据，只需要有一个能存储检索数据文件元数据的服务就能实现各个数据源之间的数据互通了。HCatalog就是这样的一个数据元数据(数据文件路径、存储格式、数据的组织格式、字段类型等)管理工具。

下面详细介绍这两个数据整合工具。

4.5.2 数据库整合工具Sqoop

Sqoop是Apache顶级项目，主要用来在Hadoop和关系数据库中传递数据。通过Sqoop，可以方便地将数据从关系数据库导入到HDFS，或者将数据从HDFS导出到关系数据库。它充分利用了MapReduce的并行特点以批处理的方式加快数据的传输，同时也借助MapReduce实现了容错。Sqoop的架构如图4-13所示。

Sqoop通过MapReduce任务来传输数据，一般只用到Map过程。Sqoop可以将HDFS中数据导入到Hive、HBase等非关系型数据库。对于关系型数据库，Sqoop通过JDBC和关系型数据库进行交互。理论上支持JDBC的数据库都可以通过Sqoop和HDFS进行交互。

Sqoop数据导入具有以下的特点。

(1) 支持文本文件(-as-textfile)、avro(-as-avrodatafile)、SequenceFiles(-as-sequencefile)。

(2) 支持数据追加，通过-apend指定。

(3) 支持table列选取(-column)，支持数据选取(-where)，和-table一起使用。

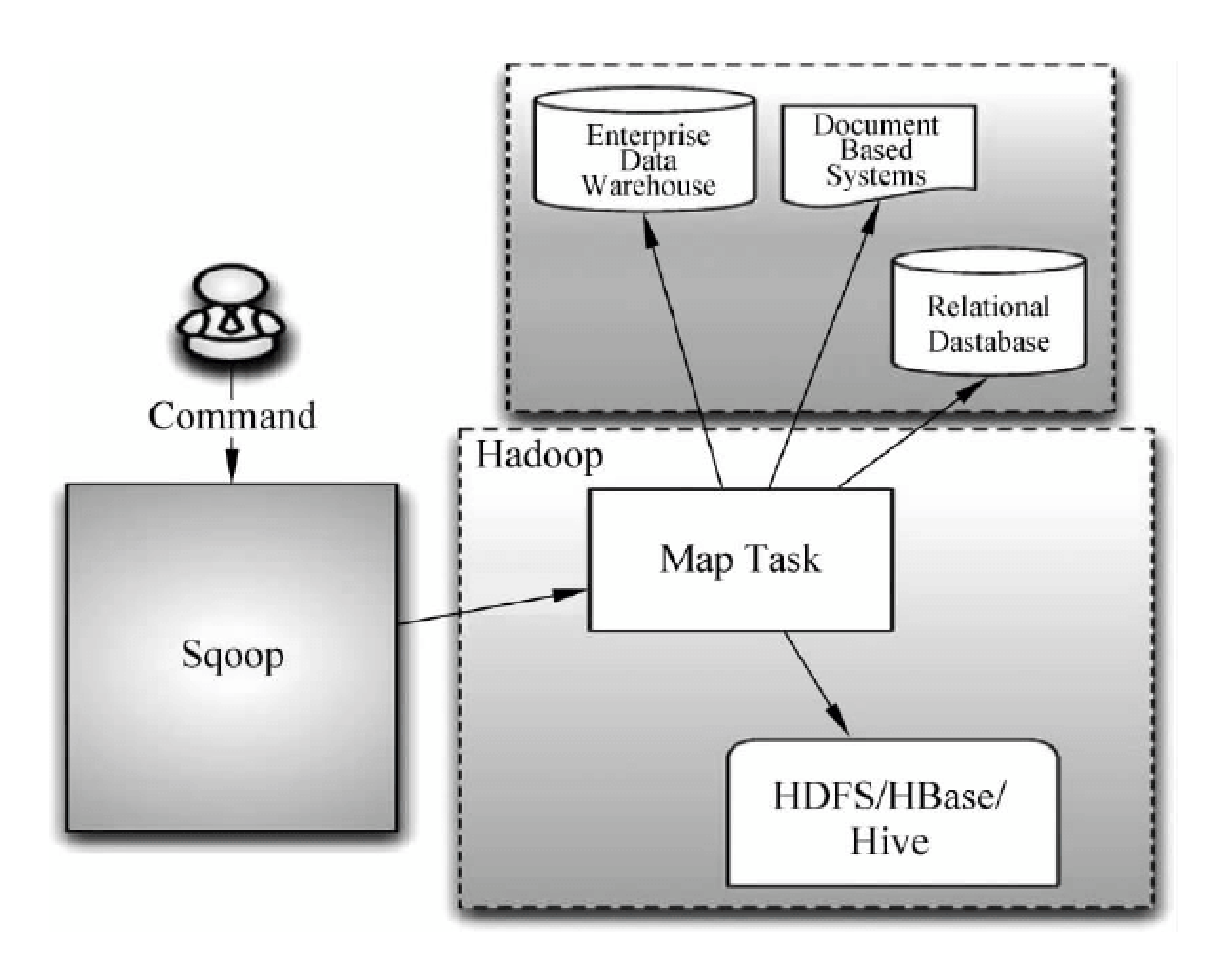


图 4-13 Sqoop 架构图

(4) 支持数据选取,例如,读入多表 join 后的数据 SELECT a. * ,b. * FROM a JOIN b on (a. id == b. id),不可以和-table 同时使用。

(5) 支持 map 数定制(-m)。

(6) 支持压缩(-compress)。

(7) 支持将关系数据库中的数据导入到 Hive(-hive-import)、HBase(-hbase-table)。

为了灵活地接入各种关系型数据库,Sqoop 将对关系型数据库的连接层抽象为一个个 Connector,从而实现了数据库连接器的插件化。因此,用户可以通过实现自己的 Connector 达到抽取自己的业务数据库数据的目的。同时,Sqoop 本身自带了许多流行关系型数据库(比如 MySQL、Oracle 等)的连接器,以方便用户的使用。

4.5.3 Hadoop 平台内部数据整合工具 HCatalog

HCatalog 是 Hadoop 的元数据和数据表的管理系统。它基于 Hive 中的元数据层,通过类似 SQL 的语言展现 Hadoop 数据的关联关系。HCatalog 允许用户通过 Hive、Pig、MapReduce 共享数据和元数据。它的另一特点就是在用户编写应用程序时,无须关心数据怎么存储,在哪里存储,还避免用户因 Schema 和存储格式的改变而受到影响。

HCatalog 应用程序的数据模型以表的形式组织,表可以放入数据库中。可以基于一个或多个键对表进行散列分区,这允许我们将包含一个(或一组)给定键值的所有行组织在一起。例如,如果使用日期对一个包含三天数据的表进行分区,那么表中将会有三个分区。可以从表中动态地创建和删除新分区。分区是多维度的,而非层次化的。分区包含多条记录。一旦创建了分区,相应的记录集就确定了,并且不能修改。记录被划分为多列,每列均有名称和数据类型。HCatalog 支持与 Hive 相同的数据类型。

HCatalog 还为"存储格式开发者"提供了一个 API,用于定义如何读取和写入保存在实际物理文件或 HBase 表中的数据(与 Hive 序列化/反序列化——SerDe 相比)。HCatalog 的默认数据格式是 RCFile。但如果数据以不同格式存储,那么用户可以实现 HCatInputStorageDriver 和 HCatOutputStorageDriver 来定义底层数据存储和应用程序记录格式之间的转换。StorageDriver 的作用域是一个分区,允许底层存储灵活地支持分区修

改，或者将不同布局的多个文件合并为一个单独的表。

以下是 HCatalog 的基本用途。

1. 实现工具间的通信

重度 Hadoop 用户绝不会使用单独的工具进行数据处理。一般情况下，用户和团队开始可能只使用一种工具，如 Hive、Pig、MapReduce 或者其他工具。随着他们对 Hadoop 使用的深入，他们会发现所使用的工具对于他们的新任务来说，不是最优的。开始使用 Hive 进行分析查询的用户，更愿意使用 Pig 为 ETL 过程处理或建立数据模型。开始使用 Pig 的用户发现，他们更想使用 Hive 进行分析型查询。尽管 Pig 和 MapReduce 这样的工具不需要元数据，但元数据的出现依然为它们带来不少益处。通过元数据存储的共享，能使用户更方便地在不同工具间共享数据。比如在 MapReduce 或 Pig 中载入数据并进行规范化，然后通过 Hive 进行分析，这样的工作流已经很普遍了。当所有这些工具都共享一个 metastore 时，各个工具的用户就能够即时访问其他工具创建的数据，而无须载入和传输的步骤。

2. 数据发现

对于大型 Hadoop 集群来说，常见的情形是应用程序和数据具有多样性。通常，一个应用程序的数据可以被其他应用程序使用，但试图发现这些情况需要大量跨应用程序的信息。在这种情况下，可以将 HCatalog 用作对任何应用程序可见的注册表。将数据在 HCatalog 中发布就可以让其他应用程序发现它们。

3. 系统集成

作为一个处理和存储数据环境来说，Hadoop 为企业应用提供了太多的机会。但为了充分使用它，必须要增强现有工具并配合使用。Hadoop 应当作为分析平台的输入，或者与业务数据存储和 Web 应用集成。组织应该享受 Hadoop 带来的价值，无须学习工具使用等新的内容。有了 Templeton 提供的 REST 服务，就可以通过常见的 API 和类 SQL 的语言将平台开放给企业。通过这种方式，它开放了整个平台。

HCatalog 在 Hadoop 集群环境中起着至关重要的作用，作为企业应用 Hadoop 的准备，HCatalog 代表着下一个合理的延伸。

4.6 大数据数据交换

工业和信息化部通信发展司副司长陈家春曾表示，中国的数据总量增长速度迅猛，预计到 2020 年将占全球的 21%，我国正向着数据资源大国的方向前进。不过，此前由于政策法规的不完善以及数据标准不统一等因素，造成我国虽然数据资源丰富，却无法实现这些资源的有效共享和应用。大数据共享交换平台的建设，将有望破解这些大数据资源瓶颈。

数据交换是指为了满足不同信息系统之间数据资源的共享需要，根据一定的原则，采取相应的技术，实现不同信息系统之间数据资源共享交换的过程。

提到数据交换，就值得介绍一个国际级的实践 EDI（Electronic Data Interchange，电子数据交换），它是一种利用计算机进行商务处理的方式。在基于互联网的电子商务广泛应用之前，曾是一种主要的电子商务模式。

EDI 是将贸易、运输、保险、银行和海关等行业的信息，用一种国际公认的标准格式，形

成结构化的事务处理的报文数据格式，通过计算机通信网络，使各有关部门、公司与企业之间进行数据交换与处理，并完成以贸易为中心的全部业务过程。EDI 包括买卖双方数据交换、企业内部数据交换等。

实际上，EDI 的发展已经至少经历了二十多年，其发展和演变的过程已经充分显示了商业领域对其重视的程度。从人们将 EDI 称为“无纸贸易”（Paperless Trade），将 EFT（电子转账）称为“无纸付款”（Paperless Payment），已经足以看出 EDI 对商业运作的影响。EDI 最初是来自于 EBDI（Electronic Business Document Exchange，电子商业单据交换）。其最基本的商业意义就在于由计算机自动生成商业单据，例如订单、发票等，然后直接通过电信网络传输到商业伙伴的计算机里。这里的商业伙伴指的是广义上的商业伙伴，它包括任何的公司、政府机构、其他商业或非商业的机构，只要这些机构与你的企业保持经常性的带有结构性的数据的交换。EDI 使用者从此项应用所得到的好处包括：节省时间、节省费用、减少错误；减少库存、改善现金流动，以及获取多方面的营销优势等。

由于实施 EDI 的最基本目的就是通过第三方服务方的增值服务，用电子数据交换代替商业纸单证的交换，而纸面单证的电子交换是建立在标准化信息基础上的，因此 EDI 的历史实际上就是商业数据的标准化和增值网络服务商的发展过程。当然，计算机之间进行电子信息传输有许多标准，特别是在不同系统的计算机之间的信息交换更是需要有很强的标准。如果排除操作系统、程序语言和其他一些硬件标准，EDI 至少涉及如下两方面的标准问题。

（1）数据标准（Data），指的是数据的格式和内容，这也是 EDI 的具体标准。

（2）协议标准（Protocol），指的是一台计算机与另一台计算机之间对话所遵循的规则。

在 EDI 的发展历史中，真正推进 EDI 发展的是那些独立的 EDI 网络增值服务商。特别是 20 世纪 80 年代以来，西方各国电信政策逐步放宽，私营网络增值服务商的出现，使 EDI 走向了商业化发展的前沿。实际上，EDI 的应用主要是来自于两个方面：一个是大的企业想与自己的供应商和客户建立电子数据交换和联系；另一个就是有些行业已经形成了非常成熟的供应链网络，通过实施 EDI 改善整个行业的整体社会效率。因此，EDI 系统较早应用在北美、欧洲、日本，以及澳大利亚的汽车制造行业、运输行业，以及日用生活用品的批发行业等。这些行业从 EDI 的应用中得到了非常好的效益。

EDI 是目前为止最为成熟和使用范围最广泛的电子商务应用系统。其根本特征在于标准的国际化，标准化是实现 EDI 的关键环节。早期的 EDI 标准，只是由贸易双方自行约定，随着使用范围的扩大，出现了行业标准和国家标准，最后形成了统一的国际标准。国际标准的出现，大大地促进了 EDI 的发展。随着 EDI 各项国际标准的推出，以及开放式 EDI 概念模型的趋于成熟，EDI 的应用领域不仅限于国际贸易领域，而且在行政管理、医疗、建筑、环境保护等各个领域得到了广泛应用。在大数据时代，数据交换的前提和目的与 EDI 都有较大的区别，然而 EDI 的实践和标准对大数据时代的数据交换仍具备很高的参考价值。

4.6.1 数据集成技术

近几十年来，科学技术的迅猛发展和信息化的推进，使得人类社会所积累的数据量已经超过了过去 5000 年的总和，数据的采集、存储、处理和传播的数量也与日俱增。企业实现数据共享，可以使更多的人更充分地使用已有数据资源，减少资料收集、数据采集等重复劳动和相应费用。但是，在实施数据共享的过程当中，由于不同用户提供的数据可能来自不同的

途径，其数据内容、数据格式和数据质量千差万别，有时甚至会遇到数据格式不能转换或数据转换格式后丢失信息等棘手问题，严重阻碍了数据在各部门和各软件系统中的流动与共享。因此，如何对数据进行有效的集成管理已成为增强企业商业竞争力的必然选择。

由于现代企业的飞速发展和企业逐渐从一个孤立节点发展成为不断与网络交换信息和进行商务事务的实体，企业数据交换也从企业内部走向了企业之间；同时，数据的不确定性和频繁变动，以及这些集成系统在实现技术和物理数据上的紧耦合关系，导致一旦应用发生变化或物理数据变动，整个体系将不得不随之修改。因此，我们进行数据集成将面临着如何适应现代社会发展的复杂需求、有效扩展应用领域、分离实现技术和应用需求、充分描述各种数据源格式以及发布和进行数据交换等问题。

在企业中，由于开发时间或开发部门的不同，往往有多个异构的、运行在不同的软硬件平台上的信息系统同时运行，这些系统的数据源彼此独立、相互封闭，使得数据难以在系统之间交流、共享和融合，从而形成了"信息孤岛"。随着信息化应用的不断深入，企业内部、企业与外部信息交互的需求日益强烈，急切需要对已有的信息进行整合，连通"信息孤岛"，共享信息。数据集成通过应用间的数据交换从而达到集成，主要解决数据的分布性和异构性的问题。数据集成是把不同来源、格式、特点、性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。在企业数据集成领域，已经有了很多成熟的框架可以利用。通常采用联邦式、基于中间件模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享和为企业提供决策支持。下面对这几种数据集成模型做一个基本的介绍。

联邦数据库系统：是由半自治数据库系统构成，相互之间分享数据，联盟各数据源之间相互提供访问接口，同时联盟数据库系统可以是集中数据库系统或分布式数据库系统及其他联邦式系统。在这种模式下又分为紧耦合和松耦合两种情况，紧耦合提供统一的访问模式，一般是静态的，在增加数据源上比较困难；而松耦合则不提供统一的接口，但可以通过统一的语言访问数据源，其中核心的是必须解决所有数据源语义上的问题。

中间件模式：通过统一的全局数据模型来访问异构的数据库、遗留系统、Web 资源等。中间件位于异构数据源系统(数据层)和应用程序(应用层)之间，向下协调各数据源系统，向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口。各数据源的应用仍然完成它们的任务，中间件系统则主要集中为异构数据源提供一个高层次检索服务。

中间件模式是比较流行的数据集成方法，它通过在中间层提供一个统一的数据逻辑视图来隐藏底层的数据细节，使得用户可以把集成数据源看为一个统一的整体。这种模型下的关键问题是如何构造这个逻辑视图并使得不同数据源之间能映射到这个中间层。

数据仓库模式：是在企业管理和决策中面向主题的、集成的、与时间相关的和不可修改的数据集合。其中，数据被归类为广义的、功能上独立的、没有重叠的主题。这几种方法在一定程度上解决了应用之间的数据共享和互通的问题，但也存在以下的异同：联邦数据库系统主要面向多个数据库系统的集成，其中数据源有可能要映射到每一个数据模式，当集成的系统很大时，对实际开发将带来巨大的困难。数据仓库技术则在另外一个层面上表达数据之间的共享，它主要是为了针对企业某个应用领域提出的一种数据集成方法，也就是在上面所提到的面向主题并为企业提供数据挖掘和决策支持的系统。

数据集成技术和方法是数据交换的基础，在大数据时代，数据交换虽然面临不同的需

求、规模、体量和应用场景，但数据集成的基础框架同样适用。

4.6.2　数据交换体系应用框架

我们总结商业公司之间的大数据交互至少有下列几种。

方式一：两家或两家以上的商业公司，他们从事的服务行业不同，拥有客户的不同方面的信息，他们的服务行业有的具有较强的相关性，整合、交互信息对其中一方或参与各方都能增加新的价值。

方式二：商业公司对社交网站的客户个人信息数据整合，期望带来新的业务增长点或实行更好的客户服务。

方式三：商业公司对政府部门的公开信息，进行大数据级别的整合和交互，产生新的商业模式、新业务，或改进客户服务。

方式四：未来，还会有新的外部大数据的整合方式会产生价值，比如某商业公司进行大量的对外部弱相关的数据的整合，当总量达到一定规模之后，仍然会产生对商业公司自身业务具有巨大价值的信息。

商业公司间的大数据种类众多，几乎大多数的情况下，两个公司之间数据的整合只对其中一方的业务有帮助，或者对双方的业务帮助价值不对等，比如社交媒体的信息对于大众商品销售公司等。因此，购买大数据的可能性远大于简单数据交换或数据互通。在这种情形下，就需要采用 4.7 节所介绍的大数据交易手段去处理了。

数据交换的体系框架包含数据源（数据交换方）、数据交换平台，以及可用于实现不同类型数据集成方案的组件和工具。我们用如图 4-14 所示的数据交换和共享平台架构来说明数据交换的体系架构。



图 4-14　数据交换共享平台架构

数据源是交换体系架构中的数据提供方，在上面的数据交换场景中，很多情形下数据提供方同时也是数据需求方，双方基于相互的数据需求在数据交换平台上进行。针对不同行业，还需要形成数据标准、元数据管理和数据字典等，才能有效地与数据交换平台进行对接。

数据接口是数据源与数据交换平台的接口，基于预先定义好的数据标准，以及数据访问格式、数据访问协议等，可以实现数据源与数据交换平台的数据对接和集成。具体的访问协议可以基于 Web Services 的模式，也可以基于其他标准的或定制化的访问协议。数据对接过程可以记录在日志中，供监控、审计、调试使用。

数据交换平台除了提供与各数据源的数据接口之外，还需要提供数据交换引擎，以及数据交换管理组件。数据交换引擎负责不同的数据模块之间的交换策略及规则；交换管理则负责相应的数据注册、认证、授权、监控等过程。

基于数据交换平台可以搭建上层的数据共享平台，比如政务共享平台、行业共享平台等，这样把下层的数据交换层形成的整合及集成数据提供给上端的数据处理及数据应用层来访问和使用。

4.6.3 数据交换关键技术

数据交换技术中最核心的技术是数据交换接口部分，一般的数据接口都是基于 Web Service(网络服务)的实现。在这个领域，有成熟的相对比较重量级的 SOAP Web Service，也有逐步成为主流的轻量级的 RESTful 服务。

1. SOAP Web Service 和 RESTful Web Service

对于 SOAP Web Service 和 RESTful Web Service 的选择问题，首先需要理解的就是 SOAP 偏向于面向活动，有严格的规范和标准，包括安全、事务等各个方面的内容，同时 SOAP 强调操作方法和操作对象的分离，有 WSDL 文件规范和 XSD 文件分别对其定义。而 REST 强调面向资源，只要我们要操作的对象可以抽象为资源即可以使用 REST 架构风格。

REST 是一种架构风格，其核心是面向资源，REST 专门针对网络应用设计和开发方式，以降低开发的复杂性，提高系统的可伸缩性。REST 提出的设计概念和准则如下。

(1) 网络上的所有事物都可以被抽象为资源。

(2) 每一个资源都有唯一的资源标识，对资源的操作不会改变这些标识。

(3) 所有的操作都是无状态的。

REST 简化开发，其架构遵循 CRUD 原则，该原则告诉我们对于资源(包括网络资源)只需要 4 种行为：创建，获取，更新和删除，就可以完成相关的操作和处理。可以通过统一资源标识符(Universal Resource Identifier，URI)来识别和定位资源，并且针对这些资源而执行的操作是通过 HTTP 规范定义的。其核心操作只有 GET、PUT、POST、DELETE。

由于 REST 强制所有的操作都必须是 stateless(无状态)的，这就没有上下文的约束，如果做分布式，集群都不需要考虑上下文和会话保持的问题，极大地提高了系统的可伸缩性。

是否使用 REST 就需要考虑资源本身的抽象和识别是否困难，如果本身就是简单的类似增删改查的业务操作，那么抽象资源就比较容易，而对于复杂的业务活动抽象资源并不是一个简单的事情。比如校验用户等级、转账、事务处理等，这些往往并不容易简单地抽象为资源。

其次，如果有严格的规范和标准定义要求，而且前期规范标准需要指导多个业务系统集成和开发的时候，SOAP 风格由于有清晰的规范标准定义是明显有优势的。我们可以在开始和实现之前就严格定义相关的接口方法和接口传输数据。

简单数据操作，无事务处理，开发和调用简单这些是使用 REST 架构风格的优势。而对于较为复杂的面向活动的服务，如果还是使用 REST，很多时候都仍然是传统的面向活动的思想通过转换工具再转换得到 REST 服务，这种使用方式是没有意义的。

SOAP 对于消息体和消息头都有定义，同时消息头的可扩展性为各种互联网的标准提供了扩展的基础，WS-* 系列就是较为成功的规范。但是也由于 SOAP 由于各种需求不断扩充其本身协议的内容，导致在 SOAP 处理方面的性能有所下降。同时在易用性方面以及学习成本上也有所增加。

REST 被人们重视，其实很大一方面也是因为其高效以及简洁易用的特性。这种高效一方面源于其面向资源接口设计以及操作抽象简化了开发者的不良设计，同时也最大限度地利用了 HTTP 最初的应用协议设计理念。同时 REST 还有一个很吸引开发者的地方就是能够很好地融合当前 Web 2.0 的很多前端技术来提高开发效率。例如，很多大型网站开放的 REST 风格的 API 都会有多种返回形式，除了传统的 XML 作为数据承载，还有 JSON、RSS、ATOM 等形式，这对很多网站前端开发人员来说就能够很好地融合各种资源信息。

REST 对于资源型服务接口来说很合适，同时特别适合对于效率要求很高，但是对于安全要求不高的场景。而 SOAP 的成熟性可以给需要提供给多开发语言的，对于安全性要求较高的接口设计带来便利。

2. 数据交换格式 XML 和 JSON

XML(Extensible Markup Language，扩展标记语言)是用于标记电子文件使其具有结构性的标记语言，可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言。XML 使用 DTD(Document Type Definition，文档类型定义)来组织数据；格式统一，跨平台和语言，早已成为业界公认的标准。

XML 是标准通用标记语言(SGML)的子集，非常适合 Web 传输。XML 提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。

JSON(JavaScript Object Notation)是一种轻量级的数据交换格式，具有良好的可读和便于快速编写的特性，可在不同平台之间进行数据交换。JSON 采用兼容性很高的、完全独立于语言文本格式，同时也具备类似于 C 语言的习惯(包括 C，C++，C#，Java，JavaScript，Perl，Python 等)体系的行为。这些特性使 JSON 成为理想的数据交换语言。

XML 的优点在于格式统一，符合标准；容易与其他系统进行远程交互，数据共享比较方便。其缺点是 XML 文件庞大，文件格式复杂，传输占带宽；服务器端和客户端都需要花费大量代码来解析 XML，导致服务器端和客户端代码变得异常复杂且不易维护；客户端不同浏览器之间解析 XML 的方式不一致，需要重复编写很多代码；服务器端和客户端解析 XML 花费较多的资源和时间。

JSON 的优点是数据格式比较简单，易于读写，格式都是压缩的，占用带宽小；易于解析，客户端 JavaScript 可以简单地进行 JSON 数据的读取；支持多种语言，包括 C、C#、Java、JavaScript、Perl、PHP、Python、Ruby 等服务器端语言，便于服务器端的解析；因为

JSON 格式能直接为服务器端代码使用，大大简化了服务器端和客户端的代码开发量，且完成任务不变，并且易于维护，因而现在 JSON 轻量级数据交换格式逐渐成为网络数据交换普遍采用的标准。

3. ESB 企业服务总线

在一些大型集团公司及企业之间，数据交换也常常基于企业服务总线（Enterprise Service Bus，ESB）的架构。ESB 是从面向服务体系架构（Service-Oriented Architecture，SOA）发展而来，是传统中间件技术与 XML、Web 服务等技术结合的产物。

ESB 提供了网络中最基本的连接中枢，是构筑企业交互系统的必要元素。ESB 采用了“总线”这样一种模式来管理和简化应用之间的复杂集成拓扑结构，以广为接受的开放标准为基础来支持应用之间在消息、事件和服务级别上动态的互连互通，是一种在松散耦合的服务和应用之间标准的集成方式。它可以作用于：

（1）面向服务的架构——分布式的应用由可重用的服务组成；

（2）面向消息的架构——应用之间通过 ESB 发送和接收消息；

（3）事件驱动的架构——应用之间异步地产生和接收消息。

ESB 的出现改变了传统的软件架构，可以提供比传统中间件产品更为低廉的解决方案，同时它还可以消除不同应用之间的技术差异，让不同的应用服务器协调运作，实现了不同服务之间的通信与整合。从功能上看，ESB 提供了事件驱动和文档导向的处理模式，以及分布式的运行管理机制，它支持基于内容的路由和过滤，具备了复杂数据的传输能力，并可以提供一系列的标准接口。ESB 用于实现企业应用不同消息和信息的准确、高效和安全传递。让不同的应用服务协调运作，实现不同服务之间的通信与整合。ESB 在不同领域具有非常广泛的用途。

（1）电信领域：ESB 能够在全方位支持电信行业 OSS 的应用整合概念，是理想的电信级应用软件承载平台。

（2）电力领域：ESB 能够在全方位支持电力行业 EMS 的数据整合概念，是理想的 SCADA 系统数据交换平台。

（3）金融领域：ESB 能够在全方位支持银企间业务处理平台的流程整合概念，是理想的 B2B 交易支撑平台。

（4）电子政务：ESB 能够在全方位支持电子政务应用软件业务基础平台、信息共享交换平台、决策分析支撑平台和政务门户的平台化实现。

4.7 大数据交易

前面一再强调，大数据的交叉融合、整合交换是充分发挥大数据价值的基础。据研究机构预测，到 2019 年，企业中 75%的分析解决方案中将需要包括 10 个或更多的数据源。其中很多的数据来自合作伙伴或第三方提供商。要整合大量的数据来源方和服务方，单纯基于数据交换的机制将很难满足复杂的业务和商务需求。就像商品市场由最初的以物换物发展到基于货币的交易市场的演进一样，大数据交易也应运而生了。

大数据交易从其市场角色和功能来说，包含但不限于以下几个方面。

（1）大数据交易可以引导对大数据商品的规范，对大数据定量、定价方面进行引导；

(2) 大数据交易应该建立认证系统，确保大数据商品的真实性和价值；

(3) 大数据交易可以深化国家有关法律对大数据商品的规范，特别是确保大数据交易的买卖双方遵守国家有关隐私、国家安全、商业机密等方面的法律，保护消费者的信息安全和其他权益；

(4) 大数据交易应该为市场参与者提供服务和手段，帮助市场参与者找到适合自己的交易方；

(5) 大数据交易应该且可以对大数据的转移和使用提供法律上的保障；

(6) 大数据交易应该且可以对大数据的转移和使用提供数据安全上的技术保障；

(7) 大数据交易应该确保资金的转移和安全；

(8) 大数据交易还可以开放大数据衍生产品，比如大数据期货，即对未来某时间段将要产生的大数据进行交易。

提到数据交易，基于美国西雅图的 BlueKai 公司可谓这个行业的先行者。BlueKai 成立于 2007 年年末，是美国著名在线数据拍卖平台，通过互联网汇集各种数据，并出售给营销人员、广告网络或内容发布商，以便增强广告质量。数据购买者中不乏全美排名前十的广告网络。BlueKai 所做的主要工作是从一些拥有部分有价值客户流量的个人或者中小网站那边购买相关信息，然后将这些信息进行分析归纳，从而总结分类出更具市场价值的流量信息，并最终进行网络拍卖。2014 年，BlueKai 被 Oracle 以 4 亿美金收购。

大数据交易在中国的发展势头迅猛，2011 年 5 月国内首家专注于互联网综合数据交易和服务的公司“数据堂”成立。2014 年 2 月，国内首个面向数据交易的产业组织——中关村大数据交易产业联盟成立，同日，中关村数海大数据交易平台启动，定位大数据的交易服务平台。2014 年 12 月，北京大数据交易服务平台上线。2015 年 4 月，贵阳大数据交易所正式挂牌运营并完成首批大数据交易。其后，在 2015 年 7 月和 10 月，东湖大数据交易所在武汉、长江大数据交易所在光谷资本大厦、徐州大数据交易所等相继挂牌成立。2015 年 12 月，河北京津冀数据交易中心成立。之后全国在北京、上海、广州、深圳、陕西、浙江等地陆续成立了一批数据交易所和交易中心，其他省市也都在规划筹建。

大数据交易所和交易中心的成立还只是大数据交易产业的开端，这个行业的黄金时期还有待时日，预计还需要至少两年的时间，这其中的原因是大数据交易还存在很多未解决的问题，还有很多基础建设需要完成，同时整个产业链的发展以及相关政策法规还需要成熟和完善。

大数据交易简单来说，首先是数据产品和数据接口的提供，然后需要进行数据资产评估，最后基于数据产品搭建数据交易平台，基于供需匹配提供数据交易服务。这其中首先需要解决的问题，就是数据的所有权、使用权、转让权的界定，以及相关的数据安全隐私考虑；其次是数据定价、评估；数据交易规则的制定；最后是数据交易平台的建设运营和相关的配套设施和服务。

从产业链条和产业生态的角度考虑，大数据交易要解决数据来源的问题。政府已经制定政策，发布了各政府部委开放数据的时间表。政府数据开放能够强化社会服务和监管，带动数据创新和产业创新。然而公开数据不构成市场交易的数据主体，还需要更多地引导和鼓励企业开放大数据。大数据的应用需要更多的企业开发各自行业、领域的数据，市场的参与者越多，市场的交易选择面和灵活度越大，能实现的价值就越大。

一旦涉及数据的开放和交易，就牵涉数据的安全和隐私。交易数据会涉及政府及行业数据、企业数据、个人数据等，除了遵守国家相关法律之外，还要设置必要的安全和隐私保护措施，对数据进行必要的脱敏处理。大数据交易还需要相关部委制定关于大数据交易的法律法规，引导市场参与者在提供大数据的同时，对于国家安全、个人隐私、商业机密等方面进行特别保护和处理。同时，大数据的交易并不局限于原始数据或是加工处理和脱敏数据的交易，还可以基于数据形成产品和服务，再进行交易，就可以有效地规避原始数据和敏感数据的隐私问题。比如通信数据和银行数据都非常敏感，直接的交易将是非法的。但如果加工成为用户的群体画像数据或是信用等级数据，就可以进行交易。

大数据作为商品进行买卖和交易，和传统的商品交易还有一定的区别。这里面涉及数据的所有权、使用权、转让权，以及数据商品的可重复使用性。一个数据包作为商品，一旦出售，是否产生所有权、使用权的转变？数据能否被购买方转卖？能不能多次重复出售？能不能同时卖给互相竞争的商家等问题，这些都跟数据属性和数据交易的细节相关。

对数据资产的评估和定价，也不是一件简单的事情。这跟数据的数量、维度、质量、性质、新鲜度、适用场景等多种属性都密切相关。一般情况下，一个数据包的价值是跟其数量和质量成正比的，包含的年度越多，用户数越多，数据的种类越丰富，精度越高，就越值钱。数据的价值，也依赖于供求关系，但同样的数据，对于不同的潜在买家，具有的价值也可能不同。所以在这些方面，也还需要探索和细化。

最后，关于大数据商品交易平台和接口的建设也是一项挑战。由于大数据的体量大、规格众多，数据如何委托、如何存放、如何交付，怎样保障数据交付的及时性、准确性，如何防范交易欺诈和风险等，都是需要认真对待的问题。

值得欣慰的是，在大数据交易领域的先行者，已经逐步积累了一些数据交易的技术、经验和规则，并在积极推动行业的进步和发展。2015 年 5 月，在 2015 贵阳国际大数据产业博览会暨全球大数据时代贵阳峰会上，贵阳大数据交易所推出了《2015 年中国大数据交易白皮书》和《贵阳大数据交易所 702 公约》，为大数据交易所的性质、目的、交易标的、信息隐私保护等指明了方向。上海大数据交易中心则发布了《数据互联规则》，在基本原则方面，强调了个人隐私保护原则、数据互联行为原则、数据权益保护原则和数据安全防控原则。个人隐私保护原则包括告知同意、选择退出、禁止公开、数据完整、维护权益、应急补救等几项原则，从各个方面维护数据主体权益。在数据互联的行为层面，发布了使用权转移原则、有限互联原则、去身份原则、负责任原则、禁止再识别原则、权利穷竭原则等，对数据交易进行了规范。另外，对于数据权益保护和数据安全防控，规则也进行了详细的条款设定。其他的交易所和交易中心也在进行有益的尝试和创新。

我们相信，随着大数据产业自身的不断壮大，对大数据品类和服务的需求也会不断增长。随着交易的进行和市场参与者的增多，大数据商品的种类会不断丰富，大数据交易的服务配套会不断完善，从而吸引更多的市场参与者，最终形成一个体量和市场巨大的新兴产业。

4.7.1 大数据交易产业链

从前面对大数据交易产业的介绍可以看到，大数据交易产业链的参与方包括数据供应方、数据需求方、平台运营方和行业监管方。当然，跟传统的商品交易市场一样，也可能会出

现数据代理方、第三方服务机构和金融机构等，在这里对他们不做过多的分析和描述。

(1) 数据供应方(卖家)：即提供某方面大数据商品的卖家，该类用户拥有某个方面的数据，通过大数据交易能形成产值和收入。

(2) 数据需求方(买家)：对相关行业数据和服务有需求的买家，购买大数据来提升自己的服务或产品。

(3) 平台运营方(平台)：大数据交易平台的运营方，通过提供大数据交易平台来收取服务费，获取利润。

(4) 行业监管方(监管)：对大数据交易进行行业引导和监管，制定行业监管相关规章制度，保障行业的平稳运营。

一个基本的交易流程包含以下几个步骤。

(1) 卖家对自己的大数据进行加工和处理，保证用于交易的大数据商品遵守国家相关的法律和规定。

(2) 卖家在交易平台上发布相关大数据商品的信息，包括数据自身属性的详细描述，也可以描述以往的交易历史，包括历史买家的行业描述等。

(3) 买家在交易平台上寻找感兴趣的大数据商品。

(4) 买卖双方就数据的使用权，数据的转移，数据是否可以再次出售(时间上，竞争对手限制等)，是否委托第三方技术公司进行数据分析等，达成协议。

(5) 买方支付交易金额，同时大数据商品转移到买方。

(6) 买方将对大数据商品进行分析或应用，实现大数据商品的价值。

大数据交易的市场参与者也可能具有多重交易身份，既是大数据的提供者，也是大数据的消费者。比如一些大数据的加工商和服务商，由于大数据商品的高价值含量，可能会先买入数据，经过处理集成后，再卖给大数据的买家。各类市场参与者的交易，能使大数据交易市场更加活跃，增加市场的流动性，带动更多的大数据商品的加入和交易。

下列公司和机构通常拥有大数据，是理想的数据供应商。

政府部门和科研机构：据统计总量数据的近 80%在政府及相关科研机构手中。比如政务、工商、税务、医疗、教育、天气、交通、道路、地质、环境以及科学研究的进展等。美国联邦政府自 2009 年就开始了政府开放数据的实践，并建立了 www.data.gov 政府数据开放网站，将 14 个大类，共十几万的数据集开放给公众，带动了全美的数据创新产业。中国政府也制定了政府部委开放数据时间表，北京、上海、杭州等地已经建立了相应的政府公开数据的网站。这些数据的开放，将极大地促进数据服务和数据交易市场的发展。

大型网络服务公司，如美国的 Google、Yahoo、微软，国内的百度、搜狐等。这类公司由于在其互联网服务领域的垄断性，累积了海量的用户和在网络行为信息。基于这些信息，他们本身就在进行相关的大数据分析和精准服务，比如搜索的相关性排名、精准广告等。另外，Google 及百度还可以利用大数据做出一些预测，如流感的爆发、政治性事件的预测、春运人群的迁徙模式等。

大型社交网站，如 Facebook、Twitter、LinkedIn、新浪微博、微信等。仅微信的全球用户就多达七亿人，每天在社交网站上产生大量的互动内容。这些网站一般都形成了自己的生态链条，通过应用开发接口，这些数据正在被大量的个人开发者和技术公司使用，用来做各种商业服务或产品推介。

大型实体商业公司或电子商务公司，如大型连锁商店 Walmart、Amazon、阿里巴巴等。这类公司大都拥有大量的客户数量、长期的客户购买记录、客户的支付历史等。这类公司最感兴趣的是客户购物的消费偏好和消费习惯。目前，这类公司的大数据应用包括推荐关联产品和推出新产品、新服务上。

大型服务公司，如银行、电信服务等公司。这类公司也拥有客户的某个方面历史消费记录，比如银行可能拥有客户的金融账户收入支出信息，电信公司拥有客户的电话或网络使用历史。这类公司通常对本行业内推出新的产品和服务，以及寻找潜在客户，降低业务风险较感兴趣。

大型制造企业，如福特汽车公司等。这类公司因为其大量的客户基础，往往可以在推出新产品服务上使用大数据技术和应用。

对于政府部委和机构的数据来说，不同部门的数据结合能够提升政府治理和管理的效率，加强对市场和行业监管，降低管理成本。对于这类数据的加工和利用将是大数据交易的一个重点板块。但同时由于政府数据存在地域性差异、数据敏感性强，还有一些涉及国家安全和个人隐私安全等情况，因此政府数据的开放将是逐步的、渐进的，针对不同部门需要形成不同的数据标准，在保障安全的基础下开放和利用。

大中型企业数据包括企业运营、管理、营销等数据以及用户信息，一些具备数据处理和分析能力的企业通过数据运营能够提升效率和服务能力，节省成本，增加营收，开拓新产品和市场。但是更多的企业仅依靠自身的数据无法实现业务闭环，在人才和资本等方面也不足以支撑企业的大数据利用，他们需要通过大数据交易平台购买相应的数据源和数据分析服务，来提升自身的数据和业务能力，这其中流通的数据也将创造更多的价值。

个体的数据可以记录和反映个人的不同兴趣、行为、意图以及偏好，依据这些数据可以对用户进行个性化的服务和精准营销。为了规避用户隐私问题，可以聚合大量个体信息形成群体信息，政府及企业可以针对群体进行分析和服务。比如美国的亚马逊网站，80%的用户再购买行为，都是基于精准推荐系统为用户所做的推荐。大数据交易平台同样也可以基于脱敏的群体和个体信息进行交易。

大数据交易平台是大数据电子交易的载体，类似于常规的商品交易平台，在平台上提供数据交易服务。数据估值由数据卖家、交易平台以及买家依据一定的规则进行协商，数据内容和交易价格在平台网站上挂出。平台则提供交易相关的支付、结算、交付及安全保障等服务。具体服务的方式以及平台所承担的功能根据 4.7.2 节大数据交易模式的讨论也会有一定的区别。

行业监管方需要保障交易的公平、开放、合规的运行，对市场交易主体、交易平台进行监管和引导，制定交易监管相应的法律法规。

4.7.2 大数据交易业务模式分析

我们总结市场上当前存在的大数据交易模式，可以分为以下三类。

1. 交易中介模式

平台仅作为一个中介方撮合买方和卖方，交易主要是数据权益的交易。平台本身不做数据存储，也不加工和分析数据。在这种交易模式下，平台只作为一个交易渠道，提供最基本的渠道(第三方中介)服务，收取渠道费，可以按次收取，也可以按月费等形式。

当前长江大数据交易所和中关村大数据交易平台都是这种交易模式。长江大数据交易所主要侧重于交易管理和交易撮合，平台自身不做数据加工分析处理。中关村数海大数据交易平台是由中关村大数据交易产业联盟发起成立，北京数海科技有限公司承建、运营。它属于开放的第三方数据网上商城，作为交易渠道，通过 API 接口形式为各类用户提供出售、购买数据(仅限数据使用权)服务，实现交易流程管理，平台按包月或调用次数进行收费。平台聚集了数千家数据供应商，数据交易额上亿元。数海还提供了必要的数据实时脱敏、清洗、审核和安全测试等基础数据处理服务。

这种模式相对比较轻量级，运营简单，完全市场化，可以借鉴传统的商品交易模式。能够有效对接供需方进行数据交易，为深层次的数据交易奠定基础。但同时由于只是提供基础的中介服务，因而还不能实现数据的深度价值挖掘和变现，平台上的数据也不能完全满足市场多样化和层次化的数据及服务需求。

2. 数据产品交易模式

这种交易模式将数据经过一定的预处理之后，打包成数据产品，然后在平台上进行出售。与上一种模式相比，平台不仅作为中介方，还需要了解买方的数据需求，要么依托自身，要么与卖方进行合作，对数据进行一定的整理、整合、清洗、脱敏、包装等处理，形成数据产品，再进行售卖。平台在数据产品的形成过程中参与度较高，与供需双方的协调合作也比较多。

这种交易模式的典型代表是数据堂。数据堂是数据交易界的先行者，成立于 2011 年，2014 年 12 月在新三板挂牌上市。数据堂主要从事互联网领域的基础数据交易和服务，自己建有交易平台。数据堂一方面提供数据产品定制模式，也就是根据需求方的要求，利用网络爬虫、众包等合法途径采集相应数据，经整理、校对、打包等处理后出售。另一方面，是与其他数据拥有者合作，通过对数据进行整合、编辑、清洗、脱敏，形成数据产品后出售。目前，数据堂拥有 4.5 万套、1200TB 以上规模的数据源，涵盖科技、信用、交通、医疗、卫生、通信、地理、质监、环境、电力等领域。

这种模式能够更好地服务于买方的需求，对数据进行定制和整合，使数据的采集、处理、交易更精准，提高了数据使用效率。但相比平台的独立性要弱一些，同时对数据的处理也是相对较基础的预处理，没有涉及深度分析和挖掘。

3. 数据再生产交易模式

这种模式比前两种模式更进一步，不局限于做大数据底层和基础数据的交易，而是提供比较深入的数据分析、挖掘、可视化服务，对数据进行再加工，再生产，将处理后的结果售卖给数据需求方。平台在交易过程中不仅提供交易服务，还提供数据存储处理对应的全链条服务，对数据进行深加工，因而能够更多地发掘数据的价值，获取更高利润。

这种交易模式以东湖大数据交易中心和贵阳大数据交易所为代表。东湖大数据交易中心是一个提供数据共享、算法服务及撮合交易的信息和技术综合服务平台，平台自身提供各项数据和分析挖掘服务；平台完全按照市场模式，以企业为运营和创新主体，整合政府公开数据、行业数据和互联网数据，打造全新的数据再生产、融合和价值发掘，运用创新金融模式，盘活政府、企业和社会的数据存量。

贵阳大数据交易则摒弃了大数据产业交易底层数据的原始概念，由交易所根据需求方

要求，对数据进行清洗、分析、建模、可视化等操作后形成处理结果再出售。数据还实行自动计价连续交易，交易所针对每一个数据品种设计自动的计价公式，数据买方可以通过交易系统查询每一类数据的实时价格。

这种交易方式更能汇聚高价值数据，包括政府部门数据和行业龙头企业数据等。同时因为交易的是数据分析结果而不是原始数据，规避了困扰数据交易的数据隐私保护和数据所有权问题，有利于活跃数据交易市场。但由于局限于交易所的数据分析挖掘能力，以及对数据的整合程度和能力，不一定能满足细分市场以及深度的行业应用的需求。

图 4-15 是国内主要大数据机构交易的一个简单的对比分析。

|主要大数据交易机构对比分析



	大数据交易所（中心）	运营厂商	交易所（中心）特点	运营厂商优势
1	贵阳大数据交易所	九次方	具有全国先发优势及标杆影响力。由于具有贵州省政府的强力支持，在政府数据公开方面具有先导作用	九次方在金融大数据行业积累了大量成功经验，在全国企业征信平台的基础上，能够为平台企业提供信用分级、风险评估、投资研究和数据服务。九次方还受邀参加工信部《中国大数据产业"十三五"发展规划》的编制工作，具有先发优势
2	长江大数据交易中心	亚信科技	以市场需求为导向，目前以企业数据交易为突破口	具有丰富的电信支撑软件提供商经验，提供覆盖电信运营商信息化运营全部环节的七百多个解决方案和三百多个软件产品。亚信本身也拥有自身的大数据产品团队，能够为企业提供大数据应用层面的服务
3	东湖大数据交易中心	中润普达	以个人数据交易为突破口，交易方式灵活	独创的中文大数据分词矩阵、信源矩阵和规则矩阵技术。公司已战略性布局数据交易领域，除了武汉，江苏、浙江、北京等地的大数据交易中心也已启动布局
4	京津冀大数据交易中心	数海科技	借助数海科技的行业经验，交易中心在数据资产评估方面具有优势，同时具有服务于京津冀的地域优势	数海科技建立了全国第一家数据交易平台，其在大数据资产评估层面拥有比较多的积累。目前正在运营的中关村数海大数据交易中心、京津冀大数据交易中心均由北京数海科技进行运营

图 4-15　国内主要大数据交易机构对比(来源-易观智库)

随着大数据交易行业的发展，交易将形成更深的行业渗透和更广的行业应用范围，可以预见到大数据的交易模式的演进，将会出现交易模式的细分，同时也会涌现出一些混合模式和混合业态。同样，就像传统的商品交易市场的演进一样，也会出现交易代理、交易中介机构，同时，基于数据产品和数据服务的衍生市场，如期货、期权等二级乃至三级市场都会逐步发展出来，我们会看到一个欣欣向荣的数据交易生态的形成。

4.7.3　大数据交易发展趋势

大数据已经成为新时代企业的资产，被类比为黄金、石油、钻石矿等高价值物品，随着大数据理论、技术、市场和应用在全球的不断深化、拓展和落地，大数据将迎来一个高速增长期，而大数据交易作为市场供需的有效媒介，也会迎来更蓬勃的发展。结合整个大数据业态的发展，我们预测大数据交易将朝着以下几个方面发展。

(1) 大数据交易的相关法律法规以及行业的标准将出台。目前这方面的实践还处在探索期，但交易一定需要法律法规的指导和监督，才能保障大数据市场的有效合规运行，而涵盖的范围包括数据标准、数据质量、数据评估、数据定价、数据安全、交易标准、平台运营、应用与服务等众多的方面。

(2) 可交易的数据类型将更加丰富。随着技术的不断进步，大数据与物联网、认知的深

度结合，将会在传感器数据、自然语言处理、语音识别、图形图像识别、影像处理、机器学习、人工智能等领域不断拓展数据的采集、加工处理和交易范围。

(3) 大数据交易的行业范围将不断扩大。随着各行业对数据的需求，以及数据开放度、数据价值变现、数据标准的不断发展，将使得更多的行业通过大数据交易的方式获得和售卖数据。

(4) 大数据交易模式和盈利模式的创新。当前大数据交易模式还比较局限，未来交易市场和交易机构会探索更加多元化、更加有效的交易模式。数据交易衍生市场也必将出现。

(5) 大数据交易机构之间的数据交易可能出现。大数据交易的真正价值在于流通，当前全国已经有很多区域性和地方性的大数据交易机构，在未来，跨区域、跨行业的数据交易和流通需求会不断增长，促进横跨大数据交易机构之间的数据交易。

(6) 需要更加关注大量长尾数据和中小企业数据应用需求。目前，交易平台的参与者主要以大企业为主，而实际上分散在众多所有者处的零散数据也非常可观，只是每个所有者所有的数据量不大，不足以让他们有意识出售数据，同时他们利用数据的观念也不强。成熟的数据交易市场需要足够数量的活跃供给方和需求方，因此未来交易平台需要激活存在于大量中小企业的长尾数据，提升中小企业的大数据应用和交易意识。

总的来说大数据交易发展前景依赖于多方面的因素，政府的支持、数据的开放、法律法规的完善、数据价值的挖掘、交易模式和方法的演进等，但其明朗的前景和辉煌的未来也是人们所共识的。



第5章 大数据管理和治理

随着大数据产业的蓬勃发展,大数据时代的商机可以说是无所不在。那么在大数据时代,提供大数据商业服务的企业的核心竞争力究竟在哪里呢?我们梳理了一个核心竞争力的三层架构,如图 5-1 所示。

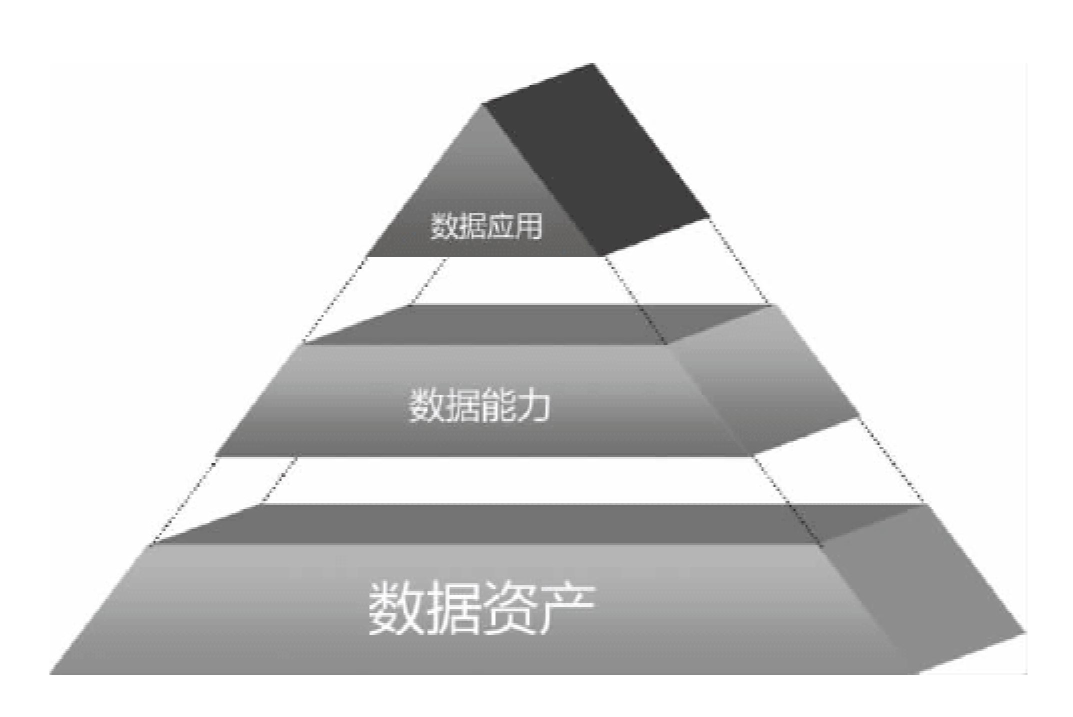


图 5-1　大数据企业的核心竞争力

由于数据已经成为新时代的资产,因而拥有数据资产的企业也就拥有了天然的竞争力,掌握着最为核心的价值。当今政府、央企所掌握的数据资源最多,随着这些数据的逐步开放,围绕这些数据所产生的商业服务,其产值也会巨大。大型互联网企业如阿里巴巴、腾讯等也掌握着庞大的数据资产,在数据产业链中居于上游地位,占据着产业优势。当然仅拥有数据资产,并不意味着立即就能数据变现,还需要具备数据处理能力。这里的能力指的是综合能力,包括技术、标准、管理、流程、人才等综合的数据处理能力。有了数据能力,就能有效地将数据资产转化为数据产品和服务,实现数据的商业价值。大型的大数据平台和工具开发商,比如 Hadoop 平台服务商 Cloudera,以及国内的如星环、清数这样的公司,拥有综合的数据处理平台和技术,能够帮助数据资源方发掘数据价值,占据了产业链的中游。在核心竞争力的末端和下游,是数据的行业应用和产业应用。即便有了数据资产,也有数据能力,开发出了数据产品,最终仍需要将产品付诸应用,才能构建一个完整的产业链条。因此,具备行业应用资源和经验,能够有效地将数据产品和行业应用结合的服务商在大数据价值链中也能占据一个重要的位置,大数据在医疗、金融、教育、工业制造、电子商务等应用领域都能发挥巨大的价值。

数据资产位居数据价值链和核心竞争力的最上游,在未来的大数据产业竞争中,将起着至关重要的作用。拥有数据资产,就拥有了制胜的先决条件,因而数据资产就是企业的命脉和根基。然而企业本身已有的数据和收集的数据并不都能成为资产。如果不能对数据进行有效的管理和治理,即使数据再多,对于企业来说也只会是垃圾和负担,非但不能成为资产,还有可能拖垮企业。传统的企业 IT 信息管理和数据管理中存在着一些比较严重的问题,阻碍着数据有效地转变为数据资产,我们总结有以下几点。

(1) 数据管理意识淡薄:没有认识到数据的重要性,没有建立数据是企业的核心资产

的意识，更多关注的是生产、销售流程等其他方面。

（2）未建立数据标准体系：大部分的传统企业都缺乏数据标准体系，数据的管理无章可循，无法可依，所依托的是散乱的人为的管理方法。

（3）不注重数据质量：一些企业即便是有数据管理的意识，也建立了一些不成体系的数据字典（元数据）或数据标准，但是数据质量参差不齐，数据没有做系统的转换、清洗、校验和结构梳理，因而不能得到有效的加工和利用。

（4）缺乏支撑数据处理的信息技术体系：传统企业由于本身信息系统的建设滞后或是不完整，更是缺乏新型数据处理的 IT 架构和体系，因而数据停留在原始的粗糙的状态。

（5）没有完整的数据生命周期管理：数据没有被作为企业信息流转的有效载体进行全生命周期的管理，只有部分过程被记录和管理，其他部分缺失，因而不能形成对数据过程的有效记录、跟踪、回溯、审计及传递，无法形成生命周期闭环。

（6）不能有效应对数据的安全和隐私问题：一旦数据的安全和隐私管理不善，所带来的后果及损失很严重，因而在数据的安全和隐私问题面前，出现了令人哭笑不得的局面：一方面知道数据蕴含着巨大的价值，一方面又不敢触碰和开放，致使很多拥有大量丰富数据资源的部委和企业，如移动、税务、银行等，谈数据开放就色变，被业界形容为“坐在金山上啃馒头”。但同时数据安全和隐私问题确实不可小觑，需要建立有效的保障体系，并采用先进的数据脱敏、泛化、加密等手段，有条件地发掘数据的价值。

以上的问题涉及大数据管理和治理的方方面面，严重阻碍了企业的数据资产化和价值变现进程。因其关系到体制、体系、标准、技术、安全等各方面，而建立和完善这些方面的过程肯定不是一天两天，在这个过程中，大量的数据在沉睡、在长灰，数据的价值被埋藏。因而快捷高效地建立好大数据管理和治理体系，是很多企业亟待解决的问题。

在数据量急剧增长，企业面临的竞争不断加剧，以及企业需要精细化运营、精准地服务客户的大趋势下，大数据管理对企业来说，不再是一种选择，而是一种必然，逃避和抵抗都是徒劳的。这反映在以下几个方面。

首先，由于企业需要处理和整合的数据源和数据量在不断增长，因而企业不得不尽快加强数据管理的基础设施和提高数据管理技能，否则，如果从基础和技术角度远远落后的话，将会很难再赶上。容量规划（产能计划）比以往任何时候都要重要，需要进行合适的调整以适应大数据指数级的增长。同样地，从商业角度来讲，延迟使用大数据将会耽误商业价值的实现和提升。

其次，企业需要将大数据融入企业数据。哪怕只是从使用一个大数据的数据仓库开始。之后需要逐步融合 Web 日志、传感器数据、运营日志等其他数据，在这个过程中，还需要判断每个类型的数据如何融入企业数据的总体架构，如何发挥价值。对企业来说，需要的是利用大数据，而不只是管理它。收集和存储大数据都要花钱，所以不要让大数据管理成为一个成本中心，需要寻找方法来从大数据中获得商业价值。当选择大数据平台来管理数据时，需要考虑成本、价值、新技术、开源技术等多方面因素。

大数据管理对企业实现商业价值体现在以下几个方面。

首先，先进的数据分析是从大数据获得商业价值的主要路径，这是很明显的事实，它甚至有一个专门的名字：大数据分析。随着大数据可用性的提高，企业对高级分析的需求也在增长，可以从其研究中获得新的商业事实和见解。

将大数据与传统数据相结合是另一种实现价值的途径。例如，对客户或其他业务实体的 360 度画像，当同时基于传统企业数据和大数据会更加完善和强大。大数据可以来自新客户接触点如移动应用、社交媒体等，就可以丰富对客户的视图。

大数据可以扩展旧的应用。这包括前面所提到的任何依赖于对用户进行全方位画像的应用。大数据也能够加强分析型应用对于数据样本的解析能力，特别是在欺诈防范，风险控制和客户细分方面。

大数据还可以催生新的应用。例如，近年来，一些货运公司和交通部门已经添加了大量传感器到每个车队的车辆和动车/高铁车体中。来自传感器的数据流使他们能够更有效地管理移动资产，更加及时可靠地运送，可以识别不符合规范的经营，并提前发现需要维修的车辆或部件，而传统的方式是很难做到的。

那么大数据的管理和治理体系，具体涉及哪些方面呢？其实上面所列出的问题，也就是我们具体要解决的方面了。我们把它们更系统地归纳为以下几个方面。

(1) 建立数据驱动的管理体系和架构；

(2) 大数据治理体系，这其中包含数据标准、数据质量、数据生命周期等关键部分；

(3) 大数据信息技术体系；

(4) 大数据安全隐私管理体系。

5.1 建立数据驱动的管理体系和架构

我们已经进入数据技术(DT)、数据驱动的时代。在企业中，传统的资产、产品、生产系统、财务系统、软件系统都有专人负责管理。那么当数据成为企业核心资产时，也需要有专门的管理机构、管理人员、管理条例等相应的管理体系和架构。企业在向数据驱动的运营模式转型和变革的时候，首先应该从组织和机构变革做起，企业应该设立专门负责数据架构和管理的组织及团队，其形式可以是实体的管理组织，也可以是虚拟的，但一定是横跨不同业务部门和项目的。这个管理组织需要不断完善数据管理和治理的架构、标准及流程，提升企业数据规划、设计、开发和交付的质量，负责数据资产的全生命周期维护，并保障数据的安全和隐私。

5.1.1 建立数据管理组织和团队

即使拥有世界上最先进的数据管理规范和指导，首先还是需要有这些规范的执行者，也就是需要首先建立数据管理组织架构和团队。结合企业自身的管理体系，在架构上一般可以分为领导决策层、部门主管层和执行层。

领导决策层：可由企业的高级管理人员来担任，负责制定企业的数据管理、数据运营、数据决策战略，并落实到中下层的具体执行策略和计划上。现在在很多大型的现代企业和从事新兴业务板块的企业中，都在设立首席数据官(Chief Data Officer，CDO)的职位。CDO 不仅是技术层面的，企业中的数据工作需要独立于业务部门、IT 部门、销售部门而存在，同时又需要和这些部门紧密相连，对业务部门、品牌部门负责。CDO 已经进入企业的最高决策层，一般直接向 CEO 汇报，可以很好地将数据的价值与企业的决策关联起来。

部门主管层可以由业务部门主管、IT 部门主管、执行项目经理等来组成和担任，也可以由专职人员来担任。很多企业还会设立专门的数据部，独立于其他部门，甚至是在企业战略

层面高于其他部门，需要其他部门来配合 CDO 和数据部制定的数据驱动战略。

在数据驱动变革方面取得很大成功的当属全球最大的职业社交平台 LinkedIn。LinkedIn 早在 2010 年就成立了独立的数据分析部门，由此部门进行的深度数据分析最后成为推动其产品、营销、服务等各部门的创新动力。很多企业只是将数据分析作为业务及 IT 部门的外延或项目管理来定位，LinkedIn 却将其作为独立部门设置，与研发、产品、市场、销售、运营等 5 大核心部门并列存在。独立的数据分析部门能够对几亿注册用户通过集成数据架构、BI、数据挖掘和分析，直接满足近 70% LinkedIn 内部员工的数据分析需求，能够覆盖和驱动其他 5 个商业职能部门。

独立的 LinkedIn 数据分析部已经几乎支撑了 LinkedIn 的所有业务，推动了 LinkedIn 主流商业模型之间的结合与相互驱动，形成一个良性增长的闭环。首先是数据分析推动了用户的增长，提高了用户的体验，其次，用户的增长和体验增加了很多后台和前台的数据，然后，LinkedIn 会从这些新的数据里面发现更多的解决方案和产品，以推动商业的增长、用户的体验和用户数量的增加，从而进入一个数据的正向循环。

具体执行层：执行层主体可能是数据部的员工，但总体数据战略的执行，会关乎企业的每一个部门和员工。正如上面提到的 LinkedIn 的案例，最终数据分析部与其他部门形成了一个良性的循环，带动了企业的全面发展。

5.1.2　建立数据管理规章和制度

很显然，在设立了 CDO 及数据部门等管理执行组织架构之后，他们所制定的数据战略需要得到有效的执行和保障，那么就需要有配套的数据管理办法、职责划分、绩效等数据管理规章和制度。这需要结合企业实际，为数据管理战略及策略的开展和执行制定切实可行的管理办法、业务流程、人员角色和岗位职责、认责体系，并建立好相应的支持环境。由于大数据管理还涉及数据管理的技术架构体系，以及大数据管理本身所用到的管理工具、管理平台、管理软件等，因而这些规章和制度还涵盖这些工具及技术的相关操作流程。

管理执行组织负责监督、管理、实施和执行与大数据管理及治理相关的一切流程与环节，包括制定并审核数据政策、标准和程序；审阅和批准数据架构；计划和发起数据管理项目和服务；评估数据资产价值和相关成本；数据管理监督和控制；监督数据专业组织和工作人员；协调数据治理活动；管理和解决数据相关问题；监控和确保遵守法律法规；监控和确保符合数据政策、标准和架构；监督数据管理项目和服务；交流和宣传数据资产的价值等等诸多方面。

5.2　大数据治理体系

数据治理指的是数据资产管理的权威性和控制性活动（规划、监视和强制执行），数据治理是对数据管理的高层计划与控制。大数据治理体系的构建为数据管理工作提供强有力的系统支撑。建立一个完整的数据治理体系可以从组织架构、标准、质量、系统功能等方面增强数据宏观管控，在微观上实现精细化管理。数据治理模块主要包括数据标准管理、数据质量管理、元数据管理、主数据管理、数据生命周期管理、数据安全管理等，这些模块协同运营，确保数据规范、一致、安全、有效。

（1）数据标准管理：建立数据标准体系，并制定数据标准运维管控制度和流程。

（2）数据质量管理：保证数据的完整性、一致性、准确性、及时性、合法性，提升用户使用体验。

（3）元数据管理：维护基础数据描述。

（4）主数据管理：管理核心数据。

（5）数据生命周期管理：重点建设从数据资产的规划、注册、运营到注销的全流程管理体系，使数据资产管理系统化、可视化。

（6）数据安全管理：建立体系化的数据安全管控策略，实现全方位数据安全管控机制，通过技术手段与管理措施相结合的方式保障数据安全。

5.2.1 数据标准管理

制定和维护数据标准对于大数据管理至关重要。如果缺乏相应的标准，那么数据管理将无章可循，数据质量也将无从保证，数据的应用、交换和共享也会混乱无序。数据标准管理体系如图5-2所示，包括数据标准的规划、数据标准的实施，以及数据标准的相关支撑。数据标准的规划包括制定数据标准体系和实施线路图；数据标准支撑部分主要是前面所提到的相关的组织架构、管理办法及制度，除此之外，还需要一些数据标准的管理工具。数据标准的实施是相对比较关键的部分，它包括标准的制定、执行、维护和监控。



图5-2 数据标准管理体系

数据标准的制定包括数据标准的编制、数据标准的审查和数据标准的发布。标准的制定需要依托一个数据标准化管理组织，该组织一般需要是一个行业性的组织，需要依托行业专家共同发起对标准的讨论、制定，修改和维护。当然也不排除制定小范围的企业内部的一些数据标准。

数据标准的编制、审查和发布过程一般有以下几个步骤。

（1）数据标准化管理组织召集数据提供者和执行者参与数据标准相关属性的收集、整理、加工等工作，并按照协商一致的原则形成数据标准初稿。

（2）数据标准初稿经过多次的讨论、修改和丰富后，形成数据标准送审稿提交给数据标准管理决策者。

(3) 经过数据标准管理决策者的讨论审核后，由数据标准化管理组织再次进行数据标准的修改完善，并完成数据标准的发布。

数据标准的执行是指数据标准的落地实施和执行过程，并且对执行过程进行监控和检查，保证标准执行到位。

数据标准的维护则是依据行业、时代和技术的发展，对标准进行必要的修订。

数据标准的监控则是对标准的执行建立考核体系，并进行日常和实施落地的监控。

从数据标准化实践来说，企业需要梳理好核心的元数据、主数据，形成相应的规范化的数据框架和模型，然后做好执行、监控和维护。这整个标准化的管理流程本身也需要规范化。

数据的标准化一般会涉及数据的编码标准，编码是用于唯一区别一条数据记录的特殊标识。编码需要统一规划、统一编制，这样可以避免各企业或企业部门各自为政，对数据进行独立的编码，导致数据整合中发生不兼容、重码、冗余、冲突等各种问题。再一个是数据的分类标准，是用于将具有相同数据属性、管理要求和系统要求的数据进行分类分组的标准。通过这样的分类标准对数据进行专项化的着重管理，并为业务管理和分析提供基础参照。数据标准还涉及数据字段和属性的规范化，即规定每个数据字段内容的填写和检验的规范，保证所有数据在整个企业或行业范围内的规则统一。数据的交互流程和业务规则也需制定相应的标准。

5.2.2 数据质量管理

数据质量可以定义为数据的"适用性"，也就是数据是否满足应用的需求，满足的程度越高，说明数据质量越高。数据质量是开发数据产品、提供数据服务、发挥大数据价值的必要前提，是数据治理的关键因素。数据质量一般需要满足准确性、完整性、一致性、及时性、合法性等多个维度。所谓准确性，就是数据必须真实准确地反映所发生的业务；完整性是指数据是充分的，任何相关的数据都没有被遗漏；一致性是指数据之间是相关的，有一定的相互约束，数据在不同场景下这种相互关联性都需要一致；及时性是数据需要及时更新，不能是过期的；合法性是指数据需要合理合法地获取和使用。

数据质量管理首先需要从管理和机制上着手，需要建立合理的数据管理机构，制定数据质量管理机制，落实人员执行责任，保障组织间高效的沟通，持续监控数据应用过程，加上强有力的督促才能保障高效优质的数据质量管理。

数据质量管理的过程包括规则制定、问题发现、质量剖析、数据清理、评估验证、持续监控等环节，同时还需结合实践进行定制和优化。首先是根据数据标准制定数据质量校验的业务和技术规则，以及对应的数据质量问题发现及管理；然后按照数据质量维度对抽样或全局数据进行剖析，并结合评估验证进行数据清理；最后通过数据质量持续监控，以数据质量报告的形式汇报并反映数据质量的状况及问题。整个过程需要形成常态化持续化的闭环，才能持续改进数据质量。数据全过程质量管理框架以改进数据质量为目标，确保数据的准确、完整、一致和及时性。

数据质量如果得不到保障，将会对业务目标的完成造成很大的影响。数据质量管理人员必须找到并使用数据质量指标，报告数据缺陷与受影响业务目标之间的关系。定义数据质量指标的过程存在着挑战，识别并管理业务相关的数据质量指标，可以与监控业务活动绩效相类比，数据质量指标应该合理地反映数据质量情况，为数据质量管理提供量化依据。在

定义数据质量指标的过程中，需要充分考虑可度量性、业务相关性、可接受程度、可控性、可追踪性等特性，并与数据认责制度充分结合。首先需要分析业务影响，并评估相关的数据元素以及数据生命周期流程；其次针对每个数据元素，列出与之相关的数据需求，并定义数据质量维度以及业务规则；最后针对业务规则，描述度量需求满足度的流程，并定义可接受程度的阈值。

数据质量问题是指数据不适合业务运行、管理与决策的程度。由于数据质量需求涉及的范围和影响程度不一，需要通过分析数据质量问题级别进行分类。较小的需求只需要对单系统数据项进行修改，处理方式相对简单；中间的需求是对业务口径、技术口径的确定；较大的需求则有大规模跨部门的系统级建设或改造需求，对其根源进行剖析甚至需要进行业务规则的调整。找到质量问题所在之后，对问题进行评估验证，并进行适当的数据清理，可以解决相应的质量问题，改善数据质量，之后进行持续的质量监控，这是整个数据质量的管理闭环过程。

5.2.3 元数据管理

元数据(Metadata)是关于数据的描述。在企业数据管理中，又可分为技术元数据和业务元数据。技术元数据是存储关于数据管理系统如数据仓库系统技术细节的数据，是用于开发和管理该数据仓库所使用的数据，它主要包括以下信息：数据仓库结构的描述，包括仓库模式、视图、维、层次结构和导出数据的定义，以及数据集市的位置和内容；业务系统、数据仓库和数据集市的体系结构和模式等。业务元数据则从业务角度描述数据仓库中的数据，它提供一个介于使用者和实际系统之间的语义层，使得不懂计算机技术的业务人员也能够“读懂”数据仓库中的数据。业务元数据主要包括以下信息：使用者的业务术语所表达的数据模型、对象名和属性名；访问数据的原则和数据的来源；系统所提供的分析方法以及公式和报表的信息。比如说企业概念模型，这是业务元数据所应提供的重要的信息，它表示企业数据模型的高层信息、整个企业的业务概念和相互关系等。

如图5-3所示为数据仓库元数据示例。

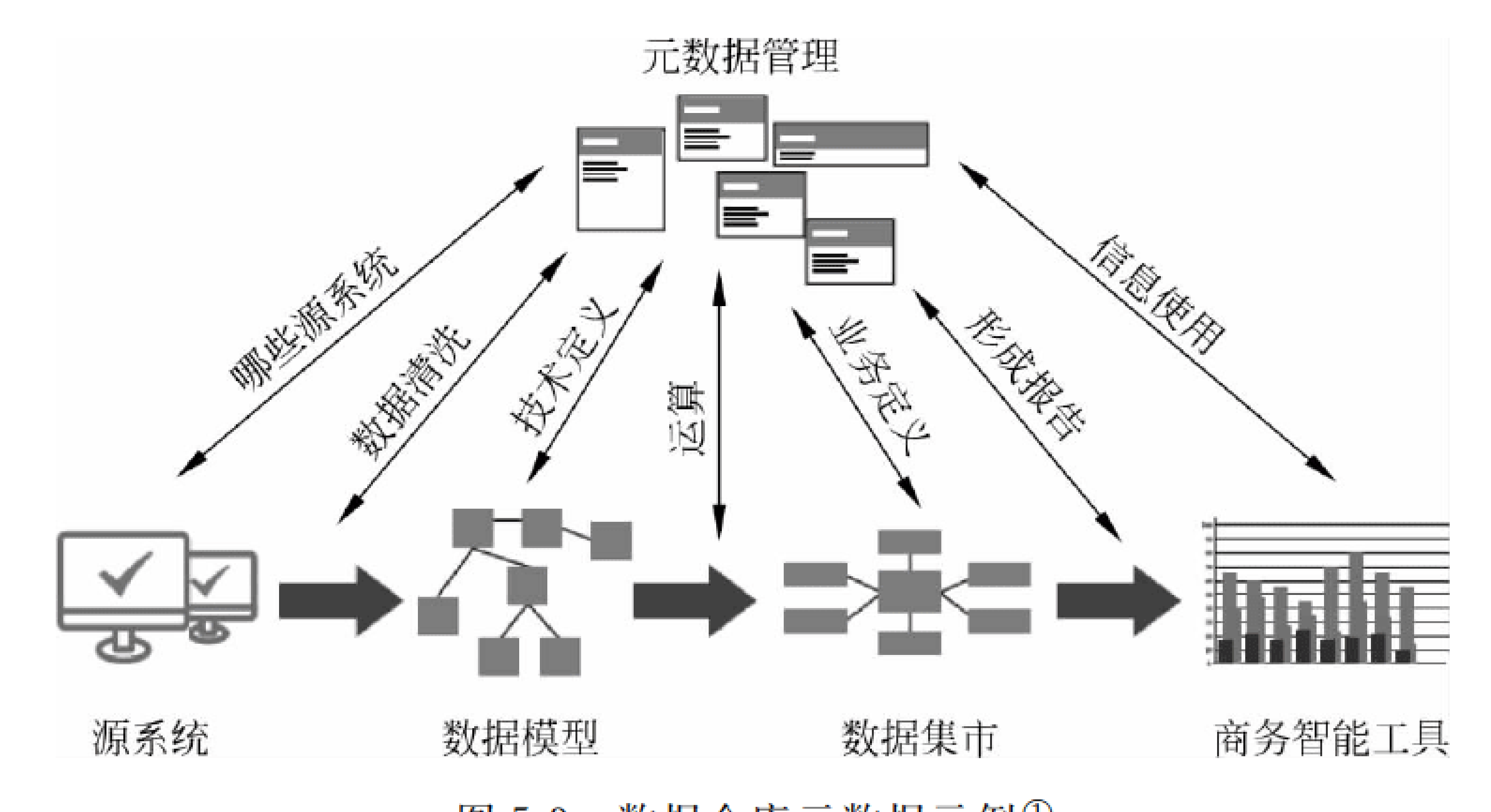


图5-3 数据仓库元数据示例①

① http://www.primeton.com/products/dg/img/how.5-1.png

元数据的管理同样包含相应的管理组织架构的设定，以及元数据管理的规章规则等，在此基础上，再定义元数据的管理流程，包括元数据的定义、元数据的变更、元数据的同步、元数据的权限管理，以及元数据检查及报告。由于元数据包含企业的基础和敏感信息，因此安全和权限管理也很重要，为加强数据安全及隐私保护，对每个元数据，可以设置对应的数据隐私级别信息；同时可以细化元数据权限，对于不同的部门、人员、角色，都只能授予对应的权限，对于权限变化要严格审批。

元数据管理的建设将贯穿大数据平台及系统的建设、使用、运营、维护的全过程，并发挥以下关键作用。

(1) 元数据提供了关键数据的详细描述，使用户了解数据组成、结构及数据流向。可以快速建立业务与技术之间的衔接，为企业管理提供重要的保障。

(2) 使用元数据管理可以自动化地获取整个企业的数据业务含义，帮助内外部客户更好地理解数据，提高数据使用的效率。

(3) 使用元数据产品能够方便内部管理、审计或外部监管的需求追溯业务指标、报表的数据来源和加工过程，追溯数据的来源。

(4) 可以追溯系统间信息生命周期，包括对数据进行的操作和流程，便于用户进行分析判断、问题定位。

(5) 元数据管理提高了信息的透明度、有效性、可访问性、一致性及可用性。它有助于依靠节约成本、提高资产价值、利益相关者满意度和卓越运营来调整 IT 投资。

5.2.4 主数据管理

主数据(Master Data)是对企业至关重要的核心业务实体的数据，比如客户、产品、订单等。这些数据分布在企业的各个业务系统之中。由于企业信息化程度的不断深入，跨业务、跨部门、跨业务系统之间的业务连贯性需求越来越迫切，因而对企业系统数据的一致性、完整性和准确性提出了新的要求。主数据是各个业务系统需要共享的数据，能帮助企业构建单一、准确、权威的数据来源。

主数据管理是制定一组规程、技术和解决方案，用于为所有跟主数据打交道的各方(如用户、应用程序、数据仓库、流程以及商业伙伴)在创建、访问和维护业务数据时，能保持一致性、完整性、相关性和精确性。

主数据管理围绕的是数据的管理，不会创建新的数据或新的垂直数据结构。它提供了规程和方法，使企业能够有效地管理存储在分布系统中的数据。主数据管理还提供先进的技术和流程，用于自动、准确、及时地分发和分析整个企业中的数据，并对数据进行验证。

主数据的管理体系如图 5-4 所示，与其他的管理体系类似，首先需要设立主数据管理的相关组织机构、管理流程及标准规范等。然后各类主数据，比如来自人事、财务、OA、ERP、CRM 等业务系统的数据，在主数据管理系统中需要注册、申请、审批、准入，然后可以进行修改和维护。当不需要或被淘汰的时候，要完成注销和废弃的过程。

5.2.5 数据资产的全生命周期管理

数据资产是指企业及组织拥有或控制的能带来经济利益的数据资源。企业的数据有可能成为资产，但不是所有数据都能具备资产的属性。数据资产包含如下几个要素：①被企

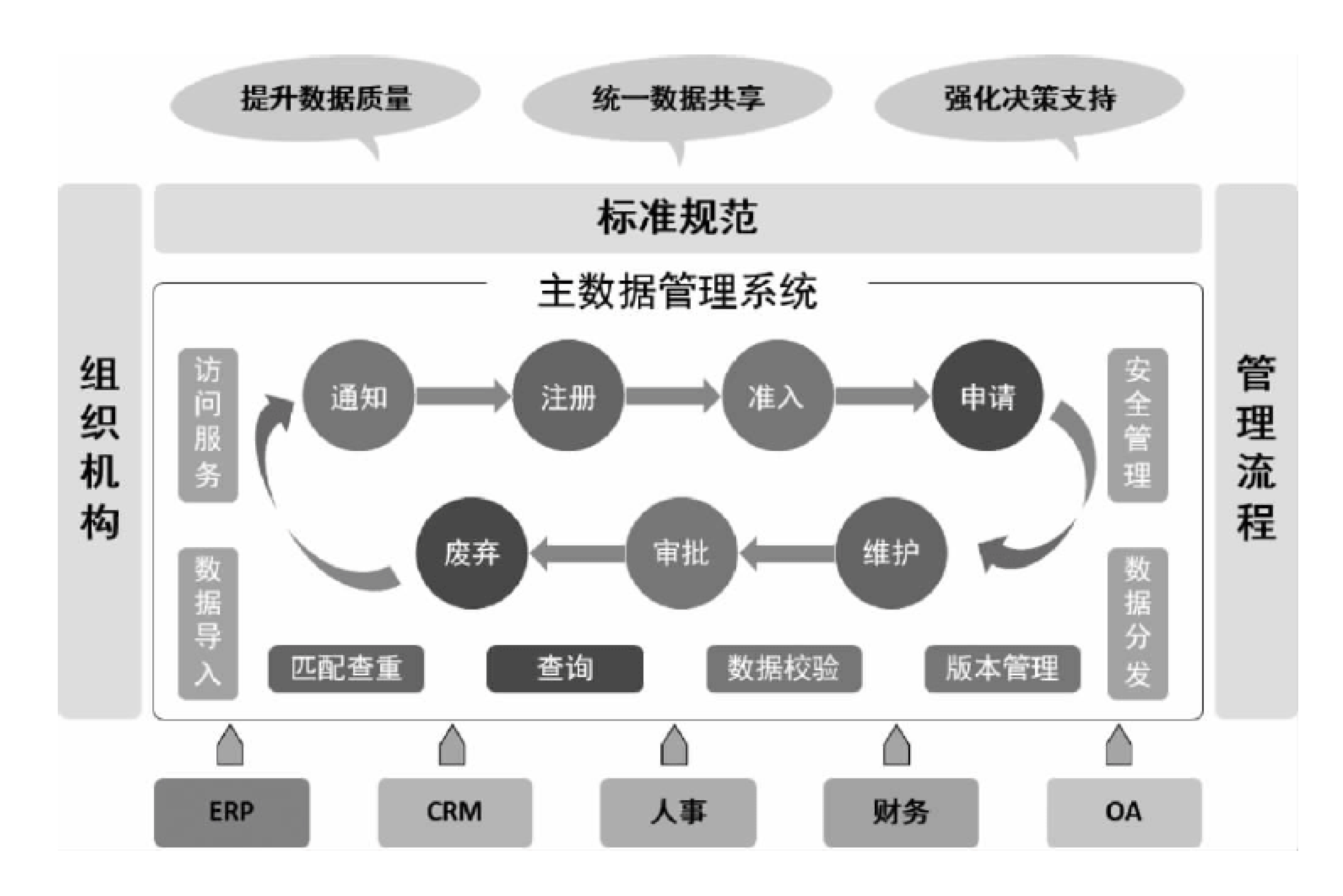


图 5-4　主数据管理体系

业拥有和控制；②能够用货币来计量；③能为企业带来经济利益。

数据资产化使得从资产的角度开展数据管理工作变为可能，将有助于多角度、全方位开展数据的管理，明确数据安全级别，落实资产责任管理，是实现数据变现的必要前提。数据资产化包含数据资产梳理盘点和数据价值评估的过程。

数据的价值根据其相关性的不同而各不相同，而数据相关性又因数据使用者而异。对某个人群没有价值的数据，可能对另外一个人群相当有用；在某个时间段内没有价值的数据，可能在另一个时间段内相当有用。

数据资产的管理如图 5-5 所示，包括 6 个部分。

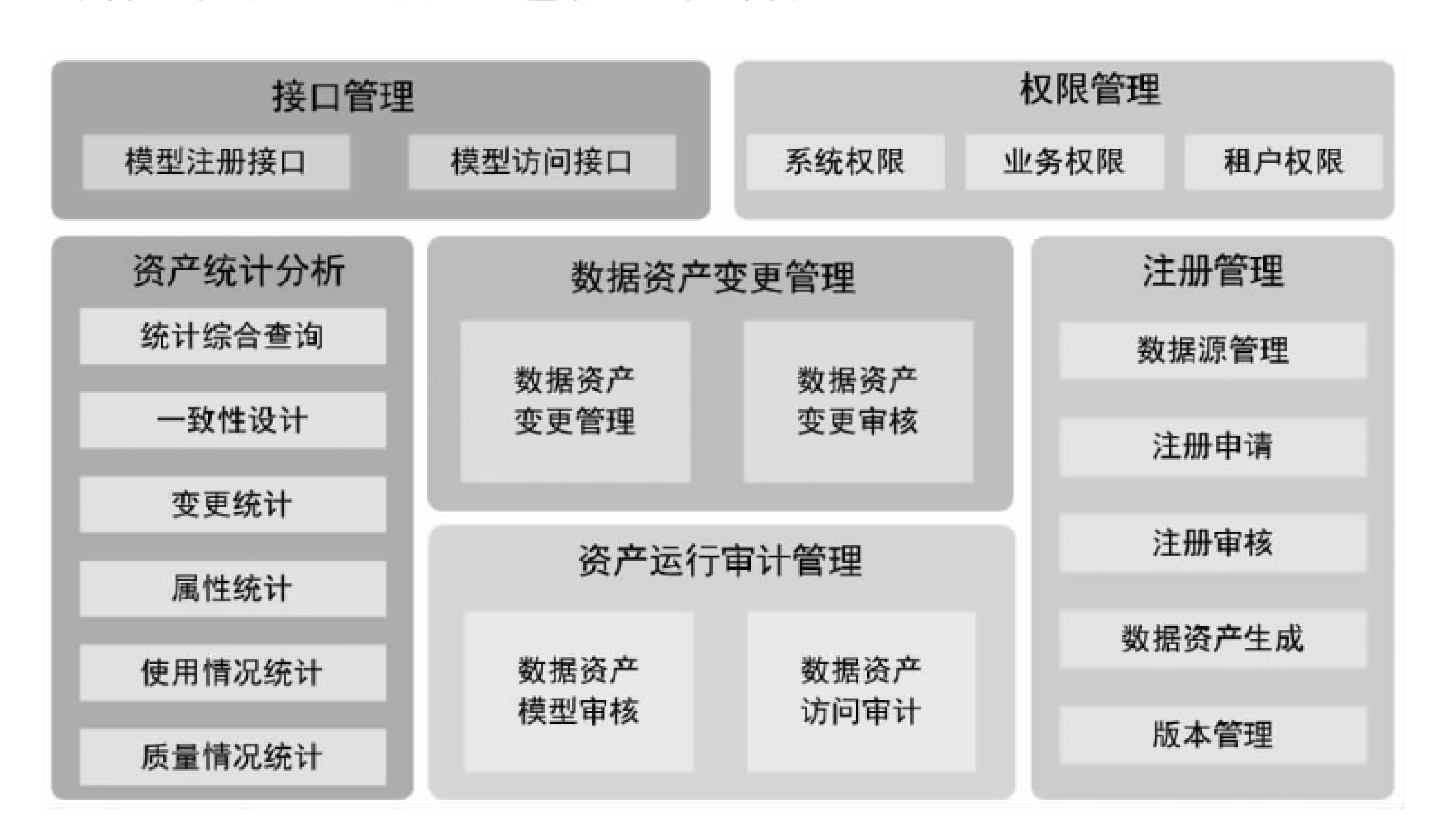


图 5-5　数据资产架构图

(1) 接口管理：与元数据管理模块、数据质量管理模块、数据安全管理模块对接，收集相关模块的基础数据，用于完成数据资产的注册、核查及安全管理等工作。

(2) 注册管理：数据资产的注册管理，并提供审核及版本控制等功能。

(3) 变更管理：支持已注册数据资产信息的变更维护，并进行相关审核。

(4) 审计管理：支持对数据资产的盘点，以及对数据资产访问记录的审计。

(5) 权限管理：对接数据安全管理模块，设置系统、业务和用户对数据资产访问的相关权限。

(6) 统计分析：支持对数据资产的属性、变更、质量、访问情况等信息的统计分析，依据这些信息还可以对数据资产进行综合评估。

定义清晰明确的数据资产信息，能有效支撑公司内部知识系统和资源管理的建设，业务人员能更快捷、有序、便利地提供资产使用的方式和途径，支撑数据分析、开发、运维的自治。数据资产化后，能实现成果和经验的共享和积累，方便实现应用和数据的生命周期的自动化管理。

数据资产管理过程是一个资产全生命周期的管理过程，以数据资产作为管理对象，以资产战略和资产策略为导向，从系统整体目标出发，统筹考虑资产的规划、投资、设计、建设、运行、维护、核查、变更、注销的全过程，在满足安全、效能的前提下有效管理与监控数据资产的生产和使用情况，不断优化数据资产质量，实现数据资产的业务价值。其管理过程如图 5-6 所示。

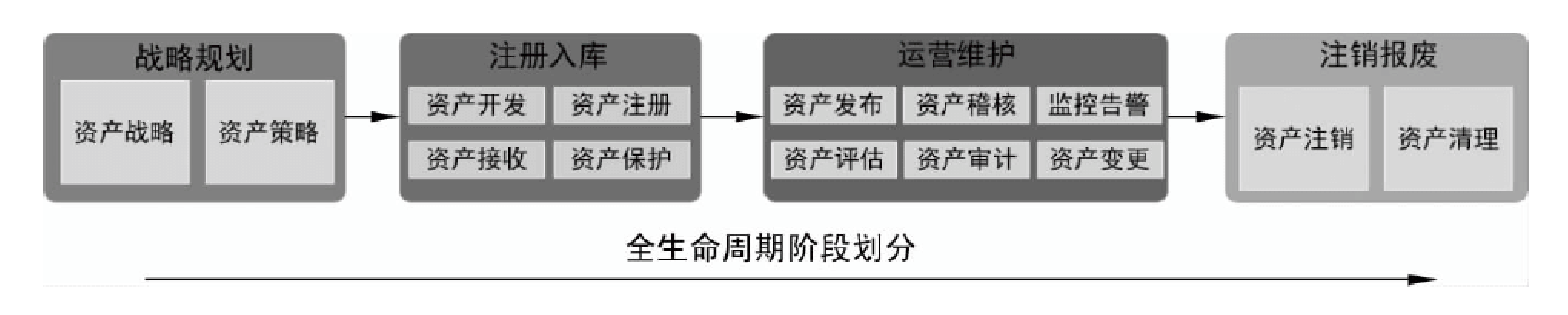


图 5-6　数据资产全生命周期管理

数据资产全生命周期管理过程分为如下 4 大阶段。

(1) 战略规划：按照业务需要和业务发展要求，建立数据资产的总体规划。制定帮助所有的数据资产供应者以及消费者运营和发展的服务战略。该阶段主要包含制定数据资产战略规划和制定数据资产策略计划等关键任务和活动。

(2) 注册入库：按照战略规划和战略计划进行数据资产的设计、建设和交付。针对需求进行分析设计，根据战略阶段的要求与规范，定义数据资产的结构等，是资产管理中的重要组成。该阶段主要包含设计和开发数据资产、数据资产注册、入库及数据资产保护等关键任务和活动。

(3) 运营维护：对数据资产的有效使用进行管控，确保数据资产健康运营。运营维护包含数据资产发布、资产稽核、监控告警、资产评估、资产审计、资产变更等方面。这些方面具体体现为：提供数据资产给授权的用户使用；对数据资产进行盘点，监控数据资产的使用情况，对数据资产访问记录进行审计；对数据资产从规划到运营阶段的情况进行全方位、多维度的统计分析，对资产内容标准化、合规性的稽核评价，根据评估结果有目的地对数据资产进行改进和完善。

(4) 注销报废：该阶段主要是对无效和失效的资产进行清理，主要包括资产注销和报废清除等任务和活动。在注销报废阶段，对已失效的资产，由管理者注销资产，并由运维者销毁资产对象。

在其生命周期中还必须建立完整的信息安全管理措施和技术方案，加强数据信息安全管控。

5.3 大数据技术管理体系

新兴的大数据正在迫使许多企业和组织进行改变。大数据的特性带来了大型数据集和非传统数据结构的许多管理困难。很多的企业正在提升他们的数据管理技能，扩展数据管理软件的投资组合，以此来提升数据管理能力。这能带给他们更多的业务流程自动化、实时化，并通过分析有价值的新的因素和属性，深入了解业务操作、流程、客户、合作伙伴等，提高运营效率和服务质量。

从大数据的技术栈来说，在第 3 章大数据平台的架构体系中详细介绍了，包括大数据的基础设施支持、采集、存储、处理、交互展示、应用以及大数据的运营管理、安全管理等。其中与传统数据管理密切相关的有数据仓库和企业 BI 商业智能，只是需要做大数据的架构升级和调整。除此之外，大数据技术管理体系还涉及大数据的流程管理、事务管理等方面。大数据的技术管理体系是和大数据的治理体系紧密结合的，帮助企业完成架构、组织、模式、标准、技术等全方位的转型和升级。

从企业的大数据管理实践角度，尽管大数据相对来说还是新兴事物，但据美国的一项调研已经有近一半的机构正在积极探索和尝试大数据管理。这其中一部分是在对现有的数据和应用的结构和关系在进行调整，打造大数据的基础数据治理地基。而另外一部分则把从 Web 服务器、设备、传感器、客户互动和社交媒体等新的数据源获得的大数据与传统数据整合在一起进行管理。

从技术实践角度，有四分之一的企业已经成功扩大现有应用和数据库来处理新兴的大数据量。另四分之一通过采用新的管理和分析多源异构的大数据专用的数据管理平台而走在最前沿。更多的则正在评估大数据平台，了解大数据的相关产品和服务，带动了大数据市场的活跃度。

据调查，Hadoop 分布式文件系统 HDFS，MapReduce 和各种 Hadoop 工具是市场上最受大数据管理欢迎的软件产品。其他包括复杂事件处理（用于数据流处理）、NoSQL 数据库（用于非结构化数据管理）、内存数据库（用于大数据的实时分析处理）、私有云等。

各个组织和机构也正在调整他们的最佳技术实践来适应大数据管理。大多数在学习 ETL——数据的抽取、转换和加载来支持数据仓库和报表。对大数据分析的准备是相似的，但有所不同。各个组织正在对现有人员进行再培训，增加顾问以增强他们的团队，招聘新的人员。重点对象是可以开发用于数据探索和发现的分析型应用的数据分析师、数据科学家和数据架构师，通过这些实践来从大数据中得到价值。

5.3.1 数据类型和结构

大数据的一个重要特点就是多样性，这就意味着数据来源极其广泛，数据类型极为繁杂。这种复杂的数据环境给大数据的处理带来了极大的挑战。而足够的数据量是企业大数据战略建设的基础，因此对多源异构数据的管理是大数据价值挖掘中的重要一环，其后的数据分析与挖掘都是建立在其基础上的。

结构化数据仍占主导。到目前为止，结构化数据是数据管理中最为常见的数据类型，占比 60%以上。这种结构化数据大多数是关系型的，这意味着关系型数据仍然是非常突出的。因而，DBMS、SQL 与其他应用于关系数据的工具类型和技术对于管理大数据仍很重要。

半结构化数据是最突出的辅助数据类型。有很多的数据混合了结构化数据、层次结构、文本等数据类型,形成了半结构化数据。常见的例子有遵守 XML、JSON 和 RSS 标准的文件。巧合的是,这些文件通常用作消息和事件的格式,因此它们也可以被认为是事件数据,其(与半结构化的数据)成为大数据的主要辅助数据类型。

网络数据。Web 服务器和 Web 应用程序已经普及二十多年,而 Web 数据又是当今大数据的常见来源。但很多的企业还没有开始管理他们的 Web 日志和点击流。Web 数据的分析和相应的 Web 优化是大数据管理的一个很好的切入点。网络数据一般都可归类到结构化和半结构化数据中。

社交媒体数据重要性在不断攀升。社交媒体网站兴起的时间并不长,用户机构近几年才开始收集社会数据进行研究。然而由于现在的年轻用户群体大部分都是在社交媒体网站上面互动,因而社交数据的管理和利用越来越重要。社交媒体由于其天然的社交属性,因而数据之间的关联是管理的重要部分,经常需要采用一些基于图的存储结构。

传感器数据,机器数据和地理空间数据开始兴起。由于物联网技术和其广泛应用,以及基于位置的服务(LBS)的兴起,很多机构正在管理和利用这些数据类型,作为比较突出的辅助数据类型。

科学数据和监测数据。一般都是科研机构和政府机构在收集和存储此类数据。

由于所有形式的非结构化数据都需要较高的专门技术和技能,因而非结构化大数据的管理具备较高的挑战性,其中比较主流的非结构化数据库的形式有人类语言或音频/视频、私人文件、电子邮件等。

5.3.2 数据存储管理

针对上述不同类型的数据,可以采取不同的大数据存储和管理方式。常用的有针对分布式文件存储类型的 HDFS 文件系统,基于此文件系统也可以进行基于 MapReduce 计算框架的分布式文档处理。Hadoop 普及程度较高,是因为 Hadoop 在管理和处理极端大数据以实现数据集成、数据仓库和分析方面的良好声誉,管理成本低。HDFS 集群已知可扩展到数千个节点,这些节点可扩展以处理数百 TB 的基于文件的数据。此外,作为数据类型不可知的文件系统,HDFS 管理非常广泛的基于文件的数据,它可以是结构化的、非结构化的、半结构化的或混合的。HDFS 集群架构与 HDFS 之上的其他 Hadoop 产品为广泛的数据密集型应用提供了一个可扩展和相对高性能的平台。

对于结构化和半结构化数据,也可采用 SQL/NoSQL/NewSQL 的存储和处理方式。如果是基于 NoSQL 的存储方式,则又需要根据具体的数据存储和处理类型进行细化的存储选型,具体的选型原则可以参考下面的一些指标。

根据不同的分类标准,NoSQL 数据库有不同的分类方式,最常用的是根据数据存储模型和特点进行的分类方式,如表 5-1 所示。

表 5-2 则列出了不同存储模型的特点和性能比较,其中,key-value 存储的操作简单,具有很高的性能、扩展性和灵活性;列存储相比灵活性要差一些,但支持的功能要相对多一些。文档存储则可以针对某些字段建立索引,能够实现关系数据库的一些功能。

上面提到的一些 NoSQL 数据库按照数据模型和查询接口分类还可以细分如表 5-3 所示。

表 5-1 NoSQL 数据库分类——按存储模型

类　　型	部 分 代 表	特　　点
列存储	HBase Cassandra Hypertable	按列存储结构。方便存储结构化和半结构化数据，方便做数据压缩，对针对某一列或某几列的查询具有 IO 优势
文档存储	MongoDB CouchDB	文档存储一般用类似 JSON(JavaScript Object Notation)的格式存储，存储的内容是文档型的，便于对某些字段建立索引，实现关系数据库的部分功能
key-value 存储	Redis Riak MemcacheDB Tokyo Cabinet Tokyo Tyrant Voldemort Scalaris Berkeley DB	可以通过 key 快速查询相应 value，不必考虑 value 的存储格式
图存储	Neo4j FlockDB	图形关系的最佳存储。如果使用关系型数据库存储的话，性能低，而且设计复杂
对象存储	Db4o Versant	通过类似面向对象语言的语法操作数据库，通过对象的方式存取数据
XML 数据库	Berkeley DB XML BaseX	高效存储 XML 数据，并支持 XML 的内部查询语法，比如 XQuery，Xpath

表 5-2 存储模型比较

	性　　能	扩 展 性	灵 活 性	复 杂 性	功　　能
关系型数据库	可变	低	低	适中	关系代数
key-value 存储	高	高	高	低	简单
列存储	高	高	适中	低	较少
文档存储	高	可变(高)	高	低	可变(低)
图数据库	可变	可变	高	高	图论

表 5-3 数据模型和查询接口

NoSQL 数据库	数 据 模 型	查询 API
HBase	列族 ColumnFamily	Thrift
Cassandra	列族 ColumnFamily	Thrift，REST
MongoDB	文档 Document	游标
CouchDB	文档 Document	Map/Reduce 视图
Riak	key-value	嵌套哈希，REST
Redis	集合 Collection	集合
Scalaris	key-value	get/put
Tokyo Cabinet	key-value	get/put
Voldemort	key-value	get/put
Neo4j	图 Graph	图

在存储多渠道来源的数据时，往往以上的一种单一存储架构并不能满足所有的存储和处理需求。这时候需要融合的存储方案，也就是底层的存储架构可能会包含文件型存储，用于存储语音、图片等数据，也会包含结构化和半结构化的数据库，用户处理交易、查询类的数据，这类实践在大数据管理领域也不少见。

5.3.3　数据仓库和商业智能

基于关系型数据库管理系统的数据仓库平台仍占据主导地位，不管是基于 SMP 还是 MPP 架构。但总的趋势明显是朝向 MPP 在发展，因为它在大规模并行数据操作方面有突出的优点。而对 SMP 来说，它仍然是操作和交易应用程序的首选架构。

在大数据时代，越来越多的企业也在采用分布式数据仓库体系结构。在多负载环境下，要设计和优化一个单一平台的数据仓库来使得所有的工作负载运行效果最佳，甚至是同时运行的时候效果最佳，是一大难题。越来越多的 DW 团队认为，一个单一的平台数据仓库不再是可取的。他们选择保留传统工作负载的核心数据仓库平台（如报告、绩效管理、OLAP），把其他工作负载卸载到其他平台上去。例如，对基于 SQL 的数据分析和处理常常卸载到 DW 设备和纵列 DBMSs。将大数据和高级分析的工作负载卸载到 HDFS、MapReduce 和其他类似的平台上。其结果是引起了分布式数据仓库架构的强劲走势。

衡量和选择数据仓库的体系结构的一种方法是计算它所支持的工作负载的数量。常规的数据仓库一般只支持最常见的工作负载，即那些标准的报告、绩效管理和在线分析处理（OLAP）。而大数据往往需要支持具有高级分析、详细的源数据和实时数据源的工作负载。因而，数据仓库的工作负载的数量和多样性的增加是企业拥抱大数据、多结构化数据、实时数据或流数据，以及用于高级分析的数据管理和处理的结果。

从企业级数据仓库（EDW）到多平台的分布式数据仓库环境（DWE）。以工作量为中心的方式导致了现今放弃单一平台的巨无霸 EDW，而转向物理分布式数据仓库环境 DWE 的趋势。一个现代的 DWE 由多个平台类型组成，包括传统的仓库和新的平台，如 DW 设备、纵列 DBMSs、NoSQL 数据库、MapReduce 工具和 HDFS。虽然多平台的方式使 DW 环境更加复杂，但是对 BI/DW 专家不是太困难的事。同时，用户可以从工作负载调整上获得高性能和高可靠的信息结果。

在操作层面，向实时的增量移动操作是当今在 BI、DW、数据管理和分析方面最具影响力的趋势。例如，实时操作（BI 和分析）需要非常新的数据在以实时或接近实时的速度收集、处理和交付。为实现这个目的，实时数据融入 EDW 已经比较常见。应用的实例包括金融交易系统、业务活动监控、效用监控、电子商务产品推荐和设施的监测监控等。

在 BI 领域，随着大数据的普及，泛 BI 的概念在大规模数据化运营的企业里正越来越深入人心。泛 BI 其实就是逐渐淡化数据分析师团队作为企业数据分析应用的唯一专业队伍的印象，让更多的业务部门也逐渐参与数据分析和数据探索，让更多业务部门的员工也逐渐掌握数据分析的技能和意识。泛 BI 其实也是数据化运营的全民参与的特征所要求的，是更高一级的数据化运营的全民参与。在这个阶段，业务部门的员工不仅要积极参与数据分析和模型的具体应用实践，更要求他们能自主自发地进行一些力所能及的数据分析和数据探索。泛 BI 概念的逐渐深入普及，向数据分析师和数据分析团队提出了新的要求，数据分析师和数据分析团队承担了向业务部门及其员工指导、传授有关数据分析和数据探索的能力

培养的工作,这正发展成为一种新兴的业态。

5.3.4 数据计算和处理

如 5.3.3 节所述,对大数据的存储管理可以采取 HDFS 分布式文件系统的方式,在其上的计算框架则一般基于 MapReduce 并行处理,通常以大规模并行处理(MPP)的形式,进行高性能数据密集型运算。MapReduce 是一个执行引擎,可以为多种编程语言编写的手工编码例程提供多线程并行性。典型的分析应用程序是在 Java、Pig、Hive 或 R 例程中编写分析逻辑,然后让 MapReduce 使用并行处理来执行例程,以访问由 HDFS 集群管理的大量的文件和数据存储库。当 MapReduce 以这种方式部署在 HDFS 上层时,结果是一个高性能分析应用程序,可以扩展到大量数据集上。

对于数据的查询、统计、分析则通常可以基于 SQL/NoSQL/NewSQL 的架构之上,对多维数据的分析和钻取则可以基于上述的数据仓库 DW/BI 的架构之上。对数据的深度分析和利用则需要依赖数据挖掘、人工智能、深度学习、社会计算等高级的处理和分析手段。

由于对数据处理的实时性要求越来越高,因而基于 Hadoop 的批处理模式在很多场景中都不足以满足性能要求,那么基于 Spark 交互式处理平台,以及类似 Storm、Spark Streaming 的流式处理平台正在成为大数据实时计算的主流平台。同时,复杂事件处理机制 CEP 作为处理多数据流的和多数据源关联的关键技术,也得到了更多的采用。流计算和 CEP 是大数据计算和处理中增长最快的技术方向。

5.3.5 数据展示与交互

数据的展示和交互技术在第 3 章也做了阐述,除了传统的二维报表和图表,还可以采用信息图(多维数据和信息的综合展示)、GIS 地图、2D/3D 图形渲染/动画,乃至可穿戴设备、可植入设备进行交互和展示。同时,随着虚拟现实/增强现实/混合现实(VR/AR/MR)在全球的普及,这些最先进的交互技术也可以被广泛应用,尤其是在大数据教育、培训、旅游、娱乐、体验等相关领域,这些技术可以发挥它们独特的优势,提供给用户很强的沉浸感和代入感。

从大数据技术的发展角度,当前出现了将探索式数据分析和可视化结合的敏捷可视化趋势。敏捷可视化允许将多种数据源结合分布式存储及内存存储,让用户进行可视化探索,包括多种可视化的组件,这些组件还可以随时增加和增强,用户可以自由灵活地对数据进行组合关联,然后选择可视化方法做近乎实时的分析和呈现。基于这样的探索,用户可以根据业务需求,快速生成业务报告,构建企业的 Dashboard(总控台或驾驶舱),并且将结果发布到企业的业务服务器上,还可以在多种设备端查看相应的分析结果并能随时做出调整。这样极大地增强了数据分析的自主性、灵活性和实时交互性。

5.4 大数据事务管理

事务是应用程序中一系列严密的操作,所有操作必须成功完成,否则在每个操作中所做的所有更改都会被撤销。事务也是并发控制的单位。事务是传统关系型数据库的逻辑工作单位,它是用户定义的一组操作序列。一个事务可以是一组 SQL 语句、一条 SQL 语句或整

个程序。

事务的开始和结束都可以由用户显式地控制,如果用户没有显式地定义事务,则由数据库系统按默认规定自动划分事务。

数据库事务特性:众所周知,关系数据库中事务的正确执行必须满足 ACID 特性,即原子性(Atomicity)、一致性(Consistency)、隔离性(Isolation)和持久性(Durability)。对于数据强一致性的严格要求使其在很多大数据场景中都无法应用。在这种情况下出现了新的 BASE 特性,即只要求满足 Basically Available(基本可用),Soft State(柔性状态)和 Eventually Consistent(最终一致)。ACID 追求一致性 C,而 BASE 更加关注可用性 A。正是在事务处理过程中对于 ACID 特性的严格要求,使得关系型数据库的可扩展性极其有限。

5.4.1 事务的基本属性

事务应该具有 4 种属性:原子性、一致性、隔离性和持久性。

1. 事务的原子性

事务的原子性保证事务包含的一组更新操作是原子不可分的,也就是说这些操作是一个整体,对数据库而言全做或者全不做,不能部分完成。这一性质即使在系统崩溃之后仍能得到保证,在系统崩溃之后将进行数据库恢复,用来恢复和撤销系统崩溃处于活动状态的事务对数据库的影响,从而保证事务的原子性。系统对磁盘上的任何实际数据的修改之前都会将修改操作信息本身的信息记录到磁盘上。当发生崩溃时,系统能根据这些操作记录当时该事务处于何种状态,以此确定是撤销该事务所做出的所有修改操作,还是将修改的操作重新执行。

2. 事务的一致性

一致性要求事务执行完成后,将数据库从一个一致状态转变到另一个一致状态。它是一种以一致性规则为基础的逻辑属性,例如在转账的操作中,各账户金额必须平衡,这一条规则对于程序员而言是一个强制的规定,由此可见,一致性与原子性是密切相关的。事务的一致性属性要求事务在并发执行的情况下事务的一致性仍然满足。它在逻辑上不是独立的,它由事务的隔离性来表示。

3. 事务的隔离性

隔离性意味着一个事务的执行不能被其他事务干扰。即一个事务内部的操作及使用的数据对并发的其他事务是隔离的,并发执行的各个事务之间不能互相干扰。它要求即使有多个事务并发执行,看上去每个成功事务像按串行调度执行一样。这一性质的另一种称法为可串行性,也就是说系统允许的任何交错操作调度等价于一个串行调度。串行调度的意思是每次调度一个事务,在一个事务的所有操作没有结束之前,另外的事务操作不能开始。由于性能原因,我们需要进行交错操作的调度,但我们也希望这些交错操作的调度的效果和某一个串行调度是一致的。DM 实现该机制是通过对事务的数据访问对象加适当的锁,从而排斥其他的事务对同一数据库对象的并发操作。

4. 事务的持久性

系统提供的持久性保证要求一旦事务提交,那么对数据库所做的修改将是持久的,无论发生何种机器和系统故障都不应该对其有任何影响。例如,自动柜员机(ATM)在向客户支

付一笔钱时，就不用担心丢失客户的取款记录。事务的持久性保证事务对数据库的影响是持久的，即使系统崩溃。正如在讲原子性时所提到的那样，系统通过做记录来提供这一保证。

5.4.2 大数据事务管理机制

通常，HBase 及 Cassandra 等 NoSQL 数据库主要是提供高可扩展性支持，在一致性和可用性方面会做相应的牺牲，在对传统的 RDBMS 的 ACID 语义、事务支持等方面存在不足。因此有很多大数据系统努力尝试把 NoSQL 与传统的关系型数据库融合，并为一致性和高可用性提供强有力的保证，我们以 Google 的 Megastore 数据库来说明大数据的事务管理机制。

Megastore 使用同步复制来达到高可用性和数据的一致性视图。Megastore 为了达到这个目标，在 RDBMS 和 NoSQL 中取了折中，将数据进行分区，每个分区进行复制，分区内部提供完全的 ACID 语义，但是分区和分区之间只保证有限的一致性。

Megastore 的底层数据存储依赖 BigTable，也就是基于 NoSQL 实现的，但是和传统的 NoSQL 不同的是，它实现了类似 RDBMS 的数据模型，同时提供数据的强一致性解决方案，并且将数据进行细颗粒度的分区（这里的分区是指在同一个数据中心，所有数据中心都有相同的分区数据），然后将数据更新在机房里进行同步复制（这个保证所有数据中心中的数据一致）。BigTable 具有一项在相同行/列中存储多个版本带有不同时间戳的数据。正是因为有这个特性，Megastore 实现了多版本并发控制 MVCC：当一个事务的多个更新实施时，写入的值会带有这个事务的时间戳。读操作会使用最后一个完全生效事务的时间戳以避免看到不完整的数据。读写操作不相互阻塞，并且读操作在写事务进行中会被隔离。

完整事务生命周期包括以下步骤。

(1) 读：获取时间戳和最后一个提交事务的日志位置。

(2) 应用逻辑：从 BigTable 读取并且聚集写操作到一个日志入口。

(3) 提交：使用分布式同步机制将日志入口加到日志中。

(4) 生效：将数据更新到 BigTable 的实体和索引中。

(5) 清理：删除不再需要的数据。

由于这类大数据的事务实现依赖于 MVCC 多版本并发控制和分区的复制机制，因而大数据的事务管理需要考虑的是跨分区的数据一致性问题，以及事务的并发性和延迟性问题。

5.5 大数据流程管理

在企业管理领域，业务流程管理（BPM）思想由来已久。在 20 世纪 90 年代，美国著名的管理学者、MIT 教授 Michael Hammer 在总结前人经验的基础上提出的“业务流程重组”和“业务流程改进”思想为现代企业全面深入进行企业流程的变革和管理奠定了坚实的理论基础。

在 IT 技术领域，业务流程管理技术的内涵也在不断地演变着，无论是侧重于人工交互的工作流系统（Workflow）厂商，还是侧重于分散系统之间整合的企业应用集成（EAI）厂商，都认为自己是业务流程产品提供商。后来 Gartner 对业务流程管理产品进行了全面的归纳和总结，提出为了实现企业端到端的流程管理，未来的 BPM 发展趋势必然是上述两类

技术的逐步融合。

企业级的流程应用，就是现在常说的建设企业端到端流程，即从客户需求端出发，到满足客户需求端去，输入端是市场，输出端也是市场。要达到这个管理目标，就对流程厂商提出了更高的要求。

首先，流程产品具备工作流特性的同时还需要具备企业级集成能力，不同的业务应用系统之间可以通过标准化的方式进行集成。为了降低 IT 建设成本，提高 IT 资产效率，还需要具备多应用系统共享同一流程产品的能力。并且，有的企业已存在多个流程产品，但并不是所有产品都具备企业级流程平台的要求，那么还需要重新规划流程平台，使得多流程平台得以共存，在复用原有流程产品成果的前提下充分发挥各流程产品特性。还有就是如果企业中存在多个开发团队，还需制定统一的流程使用接口和规范，降低使用和维护成本。

其次，由于流程产品同时具有业务含义和技术实现的特性，因此越来越多的企业 IT 部门希望业务人员能够深入参与到流程设计与开发工作中，确保流程建设过程中双方理解的一致性，降低业务部门与技术部门的沟通成本。另外，为了提升业务人员使用体验，使得业务人员无须登录多个业务系统进行流程操作，还需要规划统一用户界面访问所有流程任务。

再次，为了方便业务人员使用和维护人员维护流程，流程管理平台要能够提供统一监管视图，将有业务关联的多个流程统一管理起来，通过统一的监管视图看到业务流程的流转情况。

最后，企业为了提升自身核心竞争力，持续改进流程，需要能够建立规范的流程改进指标体系，并且通过流程平台采集这些指标数据，通过与业界标杆进行对比，发现流程中的问题，改进流程。

大数据时代的 BPM 更具备挑战性，因为需要将多源异构的数据进行整合，同时还要应对大容量、高流量、高性能要求的大数据分析处理需求。尤其是在很多行业应用中，应用程序及应用逻辑经过很多年的积累，已经完成开发，比较成熟。因此面向这类行业应用的大数据处理平台更多的是需要解决海量数据的存储和大规模计算资源及计算任务的管理调度问题。而在大部分情形下，行业应用都是由很多计算流程组成的，对这些流程的组织、分发、协调和并行化处理就成了这类行业大数据应用的关键。

基于 Hadoop 的生态体系本身也提供了 Oozie 工作流管理系统。Oozie 工作流采用 DAG(Direct Acyclic Graph，有向无环图)来定义工作流程。其中定义了一组动作(例如，Hadoop 的 Map/Reduce 作业、Pig 作业、子工作流等)，以及动作执行的顺序。图的描述采用的是 hPDL(一种 XML 流程定义语言)。

hPDL 是一种很简洁的语言，只会使用少数流程控制和动作节点。控制节点会定义执行的流程，并包含工作流的起点和终点(start、end 和 fail 节点)以及控制工作流执行路径的机制(decision、fork 和 join 节点)。动作节点是一些机制，通过它们工作流会触发执行计算或者处理任务。

Oozie 是比较高层(作业层面)的流程管理，它只是提供了一种多类型作业(比如 MR 程序、Hive、Pig 等)依赖关系表达方式，并按照这种依赖关系提交这些作业。Tez 是 Apache 最新的支持 DAG 作业的开源计算框架，它可以将多个有依赖的作业转换为一个作业从而大幅提升 DAG 作业的性能。Tez 在更底层提供了 DAG 编程接口，用户编写程序时直接采用这些接口进行程序设计，这种更底层的编程方式会带来更高的效率。

5.6 大数据易用性管理

从数据集成到数据分析，直到最后的数据解释，易用性应当贯穿整个大数据的流程。易用性的挑战突出体现在两个方面：首先大数据时代的数据大，分析更复杂，得到的结果形式更加的多样化。其复杂程度已远远超出传统的关系数据库。其次大数据已广泛渗透到人们生活的各个方面，很多行业都开始有了大数据分析的需求。但是这些行业的绝大部分从业者都不是数据分析的专家，在复杂的大数据工具面前，他们只是初级的使用者。复杂的分析过程和难以理解的分析结果限制了他们从大数据中获取知识的能力。这两个原因导致易用性成为大数据时代软件工具设计的一个巨大挑战。关于大数据易用性的研究仍处于一个起步阶段。从设计学的角度来看易用性表现为易见、易学和易用。要想达到易用性，需要关注以下三个基本原则。

(1) 可视化原则(Visibility)。可视性要求用户在见到产品时就能够大致了解其初步的使用方法，最终的结果也要能够清晰地展现出来。未来如何实现更多大数据处理方法和工具的简易化和自动化将是一个很大的挑战。除了功能设计之外，最终结果的展示也要充分体现可视化的原则。

(2) 匹配原则(Mapping)。人的认知中会利用现有的经验来考虑新的工具的使用。譬如一提到数据库，了解的人都会想到使用 SQL 来执行数据查询。在新工具的设计过程中尽可能将人们已有的经验知识考虑进去，会使得新工具非常便于使用，这就是所谓的匹配原则。如何将新的大数据处理技术和人们已习惯的处理技术和方法进行匹配将是未来大数据易用性的一个巨大挑战。这方面现在已有了些初步的研究工作。针对 MapReduce 技术缺乏类似 SQL 标准语言的弱点，研究人员开发出更高层的语言和系统如 Hive，Pig 就是一个典型的例子。

(3) 反馈原则(Feedback)。带有反馈的设计使得人们能够随时掌握自己的操作进程。进度条就是一个体现反馈原则的典例子。大数据领域关于这方面的工作较少，大数据时代很多工具其内部结构复杂，对于普通用户而言这些工具近似于黑盒子，调试过程复杂，缺少反馈性。如果未来能够在大数据的处理中大范围地引入人机交互技术，使得人们能够较完整地参与整个分析过程，会有效地提高用户的反馈感，在很大程度上提高易用性。

满足三个基本原则的设计就能够达到良好的易用性。从技术层面来看，可视化、人机交互以及数据起源技术都可以有效地提升易用性。而在这些技术的背后，元数据管理的问题是需要我们特别关注的一个问题。元数据是关于数据的数据，数据之间的关联关系以及数据本身的一些属性大都是靠元数据来表示的。可视化技术离不开元数据的支持，因为如果无法准确地表征出数据之间的关系，就无法对数据进行可视化的展示。

5.7 数据的安全管理

安全和隐私是云计算和大数据时代所面临的最为严峻的挑战！根据 IDC 的调查，安全和隐私是用户首选关注的问题，政府和企业对安全问题尤其重视，全球 51% 的首席信息官认为安全问题是部署云计算时最大的顾虑。从用户隐私角度来说，当前无论线上线下，用户的数据都收集和记录，被这些信息可能已经详细到令人很不舒服的程度。如果信息泄漏或

被滥用,就会直接侵犯到用户的隐私,对用户造成恶劣的影响,甚至带来生命财产的损失。

为了进一步明确和加强信息安全管理的规范性,可以通过制定并执行数据安全政策、策略和措施,为企业的数据和信息提供行之有效的认证、授权、访问和审计,同时还需要深化数据安全的技术防护措施。另外还需制定敏感数据访问和隐私信息保护的管理措施和技术防护措施。

数据安全主要内容包括以下几个方面。

(1) 数据权限控制,对用户的数据访问权限进行细粒度的控制管理。

(2) 客户的隐私保护,采用加密等技术手段对涉及的隐私信息进行防护。

(3) 隐私信息配置,提供隐私数据的配置服务,为隐私数据的转化服务提供识别依据。

(4) 隐私信息转化,为数据治理相关环节提供隐私信息的去隐私化或还原服务。

(5) 日志记录服务,对数据治理各环节所产生的日志记录进行收集和整理。

(6) 应用权限控制,为用户的应用功能访问权限的控制管理提供服务。

数据安全关注数据治理过程中与数据相关的安全保障技术及相应的管理办法,包括:数据权限控制、数据去隐私化、数据加解密、数据的访问记录等。数据安全为数据治理各环节提供安全保障机制及技术手段,重点关注数据治理过程中数据平台访问策略及数据资产环节的安全保障。具体的保障环节如下。

(1) 数据安全对数据平台的访问账号、功能权限进行安全保护,例如:

① 数据平台的账号管理;

② 数据平台敏感行为的控制管理;

③ 数据平台数据去隐私化。

(2) 数据安全对资产管理涉及的数据及业务过程行为进行数据安全保护,并实现相关的安全防护工作,例如:

① 数据资产的增加、删除、变更过程的数据权限控制工作;

② 数据使用过程中的防泄漏保护工作;

③ 数据资产变更过程的记录及追踪;

④ 数据粒度的权限控制管理;

⑤ 相关系统应用、数据访问行为的日志记录等工作。

数据安全需求可以分为以下几个层面。

(1) 数据存储。

① 存储设备访问控制:身份识别、权限控制、访问控制、操作审计。

② 数据安全防护:数据脱敏、数据加密。

(2) 数据处理。

数据安全防护:业务逻辑安全。

(3) 数据封装。

数据安全防护:数据最小化、数据脱敏、数据文件加水印。

(4) 数据使用。

① 接入安全控制:身份识别、权限控制、访问控制、操作日志。

② 数据安全防护:数据脱敏、数据加密、传输通道加密。

在数据安全领域,还有一种面向数据的安全体系结构(Data-Oriented Security

Architecture，DOSA），是面向数据和以数据为核心的关于数据的安全体系结构，构建起从数据保护到授权应用的整套机制。

在大数据时代，由于数据被集中存放在企业或公共的数据中心里，使信息安全问题愈发突出，急需要有一种新的安全体系结构来应对这些问题。而现有的信息安全体系是建立在相对封闭的网络环境下的，通过各种方式来保证这个封闭环境是安全的或可信的。因此，目前的信息安全，更加强调的是网络安全、系统安全、环境安全和应用安全。但是在这个相对“安全”的内部环境里，大多数数据却是处于“裸露”状态的。一旦有不速之客通过各种漏洞或非法获得权限进入这个环境，“裸露”的数据就面临着极大的危险。

一些数据中心所涉及的数据安全，多是指利用数据备份、数据灾备等技术来保障数据不丢失，但仍存在着越权访问等危险行为，造成数据和信息泄漏的隐患。

在大数据时代我们更多地面临着开放环境下的信息安全问题。随着信息系统或应用体系所面临的环境更为开放，数据共享和交换的需求越来越多，对数据和信息安全的要求也就更高。原来按照相对封闭环境下的安全举措将遇到极大的困难，不能满足新时代信息安全的要求，也给信息安全体系结构等带来了严峻的挑战。大数据时代信息安全的核心就是数据的安全，因此开展面向数据和以数据为核心的数据安全体系研究是十分必要的。

面向数据的安全体系结构 DOSA 建立在云计算基础之上，以数据“天生加密、授权使用”为原则，对数据的属性进行注册和管理，包括数据的安全属性、身份属性、时间属性、空间属性等，明确数据拥有者身份，包括数据的主人（数据权人）、朋友（被授权人）、陌生人（未授权人）和敌人（不授权人）。数据具有自保护功能，以加密方式呈现，具有不同的加密级别和深度。数据的使用要经过授权。数据是独立于系统的，数据是应用的基础，不依赖于特定的硬件环境和软件环境，同一数据可以支撑不同的应用。

面向数据的安全体系结构（DOSA）旨在从架构角度对未来的数据安全体系进行全方位设计，包括数据的管理和应用等。主要内容如下。

1. 体系结构机制及组成

包括：开放环境下数据安全的基本理论；面向数据的安全体系结构的基本原则；面向数据的安全体系结构基本构成等。

2. 数据属性

包括：数据固有安全属性；数据安全信息规范；数据状态定义及转换机制等。

3. 数据权限

包括：数据访问控制权限及管理机制；数据合法性鉴定；数据权限中心的作用和运作机制等。

4. 数据注册

包括：数据属性及数据安全信息的注册；数据注册方法；动态数据自动注册机制；数据注册信息与数据授权管理的关联机制；数据使用记录及其溯源机制等。

5. 数据授权

包括：用户认证机制及证书授权（Certificate Authority，CA）技术；用户身份与数据授权权限管理；数据授权机制及与公共密钥基础设施（Public Key Infrastructure，PKI）关系；

计账机制；多级授权及认证机制；单个数据与批量数据或大数据量授权机制等。

6. 数据加解密

包括：密钥体系；动态数据自动加密机制；数据授权自动解密机制；数据透明加解密策略和算法；加解密效率与安全性及授权过程的妥协关系等。

7. 数据应用环境

包括：传统数据传输加密技术适应性；应用环境安全保障；数据非法使用识别及数字水印技术；数据权人利益保障技术支持；数据权人权利和知识产权相关问题等。

DOSA 在组成结构方面包括数据权限中心(Data Authority Center，DAC)，数据注册中心(Data Register Center，DRC)，数据异常控制中心(Data Exception Control Center，DEC)和数据应用单元(Data Application Units，DAUs)，来实现数据的统一登记、保护、授权管理和为应用提供服务。

数据权限中心(DAC)，是 DOSA 的核心部件，对数据的安全存储、传输及应用授权进行管理。对数据实行“天生加密、授权使用”的机制，通过对数据的加解密和授权管理，使得数据在生成、存储和传输时是不可访问和使用的，而经过授权的用户在访问数据或通过应用使用数据时，是解密和透明的，即授权用户感觉不到数据的加密和解密过程。为便于管理，将数据分成存储和传输时保持加密的“数据态”和在应用中授权使用时解密的“应用态”。数据只有在“应用态”时是处于解密状态，一旦完成应用或离开了应用环境，或是由应用产生了新的数据，数据应立即“变”为加密的“数据态”，充分保证数据的安全及使用的授权。“数据态”的数据，既适合于封闭环境，也适合于开放环境，而“应用态”的数据，仅适合于“封闭”环境。数据的访问和应用是基于授权的，特定的访问者，特定的场合(环境)，特定的时间(时段)，数据的使用和用户适合于网络安全的授权、认证和计账(Authorization，Authentication，Accounting，AAA)机制。

数据注册中心(DRC)，是 DOSA 的关键部件，注册有关数据的各种信息，包括安全属性信息，通过它来构建逻辑的数据资源池，并管理数据和提供数据服务。

数据异常控制中心(DEC)，是 DOSA 的重要部件，对数据资源进行自适应管理，保证数据的唯一性和一致性。

数据应用单元(DAUs)，是 DOSA 的关键部件，关联应用对数据的访问，对各种应用提供支持。

DOSA 作为一种数据安全理念和机制，就是要保证数据能够在数据和应用两个层面中都能做到安全、可靠以及便于管理和使用，既可以在传统的封闭环境下应用，增强数据的安全保护，又可以在开放环境下保护数据的安全和不被越权访问。

目前有关信息安全、数据安全的理论和方法体系，有关网络授权、认证和计账的 AAA 技术，有关 CA 技术、PKI 技术、密钥体系、加解密技术，有关可信技术，以及不断发展的网络空间安全技术、系统安全技术、应用环境安全技术等，都能在 DOSA 框架下使用，但需要进一步从面向数据和以数据为核心的角度，进行重新梳理，从数据安全的理念、理论、方法和受保护数据的应用机制等方面，进行适应性和深入的研究，为进一步提高信息安全提供保障。

关于数据隐私保护，其技术效果可用“披露风险”来度量。披露风险表示攻击者根据所发布的数据和其他相关的背景知识，能够披露隐私的概率。那么隐私保护的目的就是尽可

能降低披露风险。隐私保护技术大致可以分为以下几类。

(1) 基于数据失真(Distortion)的技术。数据失真技术简单来说就是对原始数据“掺沙子”,让敏感的数据不容易被识别出来,但沙子也不能掺得太多,否则就会改变数据的性质。攻击者通过发布的失真数据不能还原出真实的原始数据,但同时失真后的数据仍然保持某些性质不变。比如对原始数据加入随机噪声,可以实现对真实数据的隐藏。当前,基于数据失真的隐私保护技术包括随机化、阻塞(Blocking)、交换、凝聚(Condensation)等。例如,随机化中的随机扰动技术可以在不暴露原始数据的情况下进行多种数据挖掘操作。由于通过扰动数据重构后的数据分布几乎等同于原始数据的分布,因此利用重构数据的分布进行决策树分类器训练后,得到的决策树能很好地对数据进行分类。而在关联规则挖掘中,可以在原始数据中加入很多虚假的购物信息,以保护用户的购物隐私,但同时又不影响最终的关联分析结果。

(2) 基于数据加密的技术。在分布式环境下实现隐私保护要解决的首要问题是通信的安全性,而加密技术正好满足了这一需求,因此基于数据加密的隐私保护技术多用于分布式应用中,如分布式数据挖掘、分布式安全查询、几何计算、科学计算等。在分布式环境下,具体应用通常会依赖于数据的存储模式和站点(Site)的可信度及其行为。

对数据加密可以起到有效地保护数据的作用,但就像把东西锁在箱子里,别人拿不到,自己要用也很不方便。如果在加密的同时还想从加密之后的数据中获取有效的信息,应该怎么办?最近在“隐私同态”或“同态加密”领域取得的突破可以解决这一问题。同态加密是一种加密形式,它允许人们对密文进行特定的代数运算,得到的仍然是加密的结果,与对明文进行运算后加密一样。这项技术使得人们可以在加密的数据中进行诸如检索、比较等操作,得出正确的结果,而在整个处理过程中无须对数据进行解密。比如,医疗机构可以把病人的医疗记录数据加密后发给计算服务提供商,服务商不用对数据解密就可以对数据进行处理,处理完的结果仍以加密形式发送给客户,客户在自己的系统上才能进行解密,看到真实的结果。但目前这种技术还处在初始阶段,所支持的计算方式非常有限,同时处理的时间开销也比较大。

(3) 基于限制发布的技术。限制发布也就是有选择地发布原始数据、不发布或发布精度较低的敏感数据,实现隐私保护。这类技术的研究主要集中于“数据匿名化”,就是在隐私披露风险和数据精度间进行折中,有选择地发布敏感数据或可能披露敏感数据的信息,但保证对敏感数据及隐私的披露风险在可容忍范围内。数据匿名化研究主要集中在两个方面:一是研究设计更好的匿名化原则,使遵循此原则发布的数据既能很好地保护隐私,又具有较大的利用价值;二是针对特定匿名化原则设计更“高效”的匿名化算法。数据匿名化一般采用两种基本操作:一是抑制,抑制某数据项,亦即不发布该数据项,比如隐私数据中有的可以显性标识一个人的姓名、身份证号等信息;二是泛化,泛化是对数据进行更概括、抽象的描述。譬如,将年龄 3 泛化为[0,5],把详细住址泛化为某个城区或乡镇等,可以降低信息的精确性,起到一定的隐私保护作用。

另外,从隐私保护的管理保障角度来说,可以采取三权分立的管控制度。三权是指:数据管理权限、隐私数据安全管理权限以及审计权限。三个权限分别掌握在不同的管理员手上,三个管理角色的权限相互独立、互不重叠,不允许越权,且相互制衡。

数据管理员角色:数据管理员主要负责数据平台的维护和管理,数据库设计方案及规

划。拥有数据最高的操作权限。经过隐私保护实施后，数据库中将不包含任何隐私信息。该角色能够获取所有的数据但无法读懂隐私信息，他无法获取隐私信息保护的策略和密钥信息。

安全管理员角色：是隐私数据保护专用管理角色，主要负责获取隐私信息属性，管理和配置去隐私处理的策略和密钥信息，制定版本更新计划和历史版本归档工作。该角色掌握所有去隐私处理使用的策略和密钥，但没有访问任何主数据库的权限，也无法获取隐私信息。

审计专员角色：属于专门的事后审计管理角色，审计专员有权限对数据管理员和安全管理员的任何操作进行审计。一旦发现违规的行为可以及时通告和升级处理。

建立三权分立管控制度的目的就是要建立权力制衡的机制，进一步保证隐私信息的安全。在实施过程中，必须要明确三个角色权限由不同的人员担任，三个角色的权限不能有任何的设置重叠，需配套建立相应版本更新、数据需求、后台运维、日志审计管理流程。

从全局来说，大数据管理与治理的目的是为了安全有效地管理数据，建立数据全流程的管理组织架构、管理措施、管理对象、标准、策略、技术方法、安全及隐私机制等，它是深入利用数据，发掘数据价值的基础。随着整个社会对数据价值的认识不断深化，大数据管理和治理的重要性也必将被政府、企业及行业提升到前所未有的高度，其管理与实践也会随着大数据的产业发展而更新和进步。



大数据创新方法论

6.1 大数据的爆发

从 20 世纪 80 年代到 20 世纪 90 年代，就已经有人提出数据爆炸的概念，那为什么近些年大数据才迅猛爆发呢？有 4 个方面的原因。首先，是各种各样数据源的出现和爆发。国内外大型互联网公司如 Google、Facebook、Twitter、腾讯、百度、阿里巴巴等每分每秒都在产生数据。就拿微信来举例，全球有 7 亿多用户，每个用户至少都在 10 个群里，如果每天早上一个群里发出 10 条早安、早上好的信息，那么一个早上就会产生 700 亿条信息，更不用提里面产生的各种语音和视频信息了。那么腾讯的数据中心需要实时处理这些数以亿计的信息，其后台有数百万的服务器，腾讯在内蒙古等地新建的数据中心服务器的数量在 20 万台以上。另一方面，物联网传感器、智能设备、移动终端的数量也在呈指数级增长，全球的传感器、移动终端数目都多达几百亿，智能电表也有上亿，这些设备也在时时刻刻产生和传送着数据。另外一个数据源是科学仪器和医疗仪器等。欧洲核子物理中心牵头建造的大型强子对撞机 LHC 一年就产生 15PB 的数据，通过对其中的数据进行分析，科学家们成功地发现了上帝粒子 Higgs Boson。建造在美国新墨西哥州的大型天文望远镜，旨在对地球上空四分之一的太空进行拍照扫描，记录几十亿星体的相关数据。新型的医疗仪器如高分辨率的 CT，核磁共振仪，基于直线加速器的癌症治疗设备等，基于数百万病人所产生的数据也很庞大。业界的一些概念如数据爆炸、数据暴雨、数据海啸，都是对这种数据迅猛增长的趋势的叫法。这些不同种类的数据源所产生的海量数据，正在将我们淹没，我们缺乏有效的存储、处理这些数据的手段，对数据巨大潜能的利用才刚刚开始。

微软公司有两位全球知名的科学家，一位是微软研究院副总裁 Tony Hey，他主要负责微软研究院与全球高校的合作，以及微软的交叉学科以及科学计算的研究，另一位是图灵奖获得者 Jim Gray。他们对大数据及数据科学的发展做出了巨大贡献。Tony Hey 指出当前在数据暴雨时代，虽然数据蕴藏着巨大价值，但由于数据管理以及数据技术的局限，政府、企业以及社会对数据的利用率还不到 5%，其余的数据，全都像雨水一样，通过下水道流走了。Jim Gray 由于发明了数据库的“事务”机制，奠定了全球金融交易的基石，获得了计算机界的最高荣誉“图灵奖”——相当于其他行业的诺贝尔奖。这两位科学家早在 2009 年，就提出数据密集科学，是科学的第四象限，把数据科学和之前的实验科学、理论科学、计算科学分离出来，形成一门新的独立的学科，从而带动了数据科学的研究和发展。

大数据爆发的第二个原因，是由于数据的种类、格式多种多样，数据分析的复杂度越来越高。我们举一个癌症治疗里靶向药物的例子（如图 6-1 所示）。当今癌症治疗里最为先进

的方法是使用靶向药物，它们能精准定位癌细胞，与癌细胞相结合，并摧毁或抑制癌细胞的生长。然而为了找到有效的靶向药物，需要把癌细胞和数百万蛋白质进行比对，计算的任务数多达 400 多万个，计算量在单个 CPU 上需要 50 年。也就是说，来了一个病人，我们告诉病人要等待 50 年才能把药物计算出来，这是不可接受的。

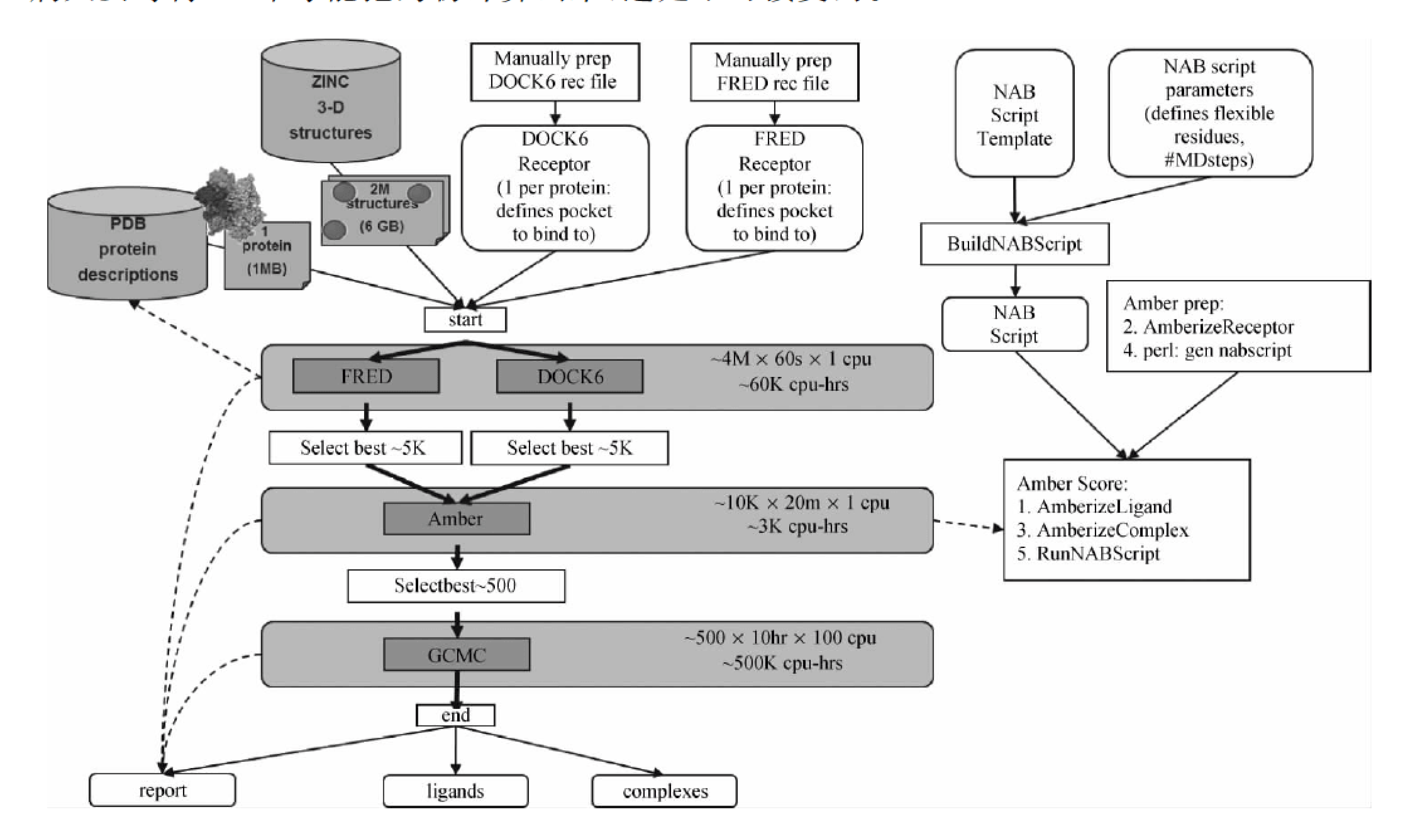


图 6-1　癌症靶点药物的计算

大数据爆发的第三个原因，是数据价值的凸显。以前，数据只是被简单地用来进行一些统计分析，甚至是放在文件柜里接灰：很多银行、医院、商家累积了几十年的数据，但都没有发挥它们应有的价值。当今，大数据被誉为新时代的黄金和石油，李克强总理称之为“钻石矿”。后面我们会看到，大数据甚至比这些还要值钱。政府、企业、社会和每个个人，都认识到了数据中蕴藏的巨大价值。从国家发展战略上，它关乎一个国家的全球战略经济布局、国计民生、政策法规、行业监管等方面；从区域经济发展来看，它是制定区域规划、城市发展计划，把握先机，占领制高地的基础；对于企业发展，它则是企业制定市场策略，规划发展和投资前景，把握先机，维持竞争力的必要手段；对个人来说，它则是工作就业、居家生活、投资理财、旅游出行等各方面的好帮手。正是由于大数据的这些核心价值，需要激活和发掘，引发了大数据的崛起。

大数据爆发的最后一个原因，也是最根本的原因，是现有信息系统已经面临严重的局限，完全不能处理数据的迅猛增长和对数据价值挖掘的渴求的矛盾冲突，从而全面引发了大数据问题在各行各业的爆发。传统信息系统的局限有以下几方面。

（1）速度方面的问题：大数据从存储到处理到展示都需要快速实时，现有系统面临瓶颈。

（2）种类和架构问题：传统系统更擅长处理结构化数据，而不是多源复杂格式数据的存储和处理。

（3）体量及扩展性问题：传统集中式处理方式无法应对海量的数据，需要分布式可扩

展的架构和系统。

(4) 成本问题：新型的分布式架构对比传统的大型计算机、小型计算机的软/硬件及运营维护成本都大大节约。

(5) 价值挖掘问题：如何从数据中挖掘出价值，并且有好的投入产出。

(6) 安全及隐私问题：在充分发挥数据价值的同时如何保障数据的安全和隐私。

(7) 互连互通和数据共享问题：打通不同机构、行业的数据互通，实现共享。

既然传统信息系统面临诸多局限，那么面对大数据的挑战，我们该如何应对，如何处理海量、多源、多结构、高流量、高通量的数据，并且有效地发掘和利用其巨大的价值和潜能呢？这需要依赖我们称之为新一代信息系统的现代科技。新一代信息系统有4大核心系统，也称为4架马车，即：云计算、大数据、物联网、移动互联网。相信读者都听说过这些概念和名词，但它们具体的含义和关系是怎样的呢？我们用一个通俗易懂的比喻来解释：如果把新一代信息系统比作一个人的话，那么物联网相当于人的眼睛、鼻子、耳朵、手等感官，可以感知周围的世界并采集数据，比如周围的温度、湿度，物体的材质等，这也是为什么物联网的采集设备被称为传感器的原因之一；移动互联网相当于神经和传导系统，可以把感官感知的数据传达回大脑；云计算则相当于身体和心脏，为思考和加工数据提供必要的能量；大数据则是新一代信息系统最核心的组成，它相当于我们智慧的大脑，只有经过大脑的加工，才能把数据转化为知识和智慧，也才能指导我们进行决策和行动。在我国，云计算、物联网都经历了很多年的发展，但是都没有取得根本性的突破和大范围的应用，最主要的原因是大数据还没有发展起来，而现在随着大数据的兴起，迅速带动了云计算及物联网的发展。国内的云计算服务商每年都有成倍的增长，而物联网又重新抬头，被称为下一个万亿级的市场，这些都是大数据技术和应用在逐步落地带动起来的。

上面总结了4个方面的原因，是大数据在近年来迅猛爆发并横扫全球的主要原因。2011年，麦肯锡在其发布的白皮书《大数据的下一个前沿：创新、竞争和生产力》中，正式提出了大数据的概念。2012年，美国奥巴马政府发布了《大数据发展和研究倡议》，把大数据列为美国的国家战略，并拨付两亿美元专款支持大数据，从而带动了大数据在全球发展的浪潮。2016年3月，我国在国家的“十三五”规划中，也正式将大数据列为国家发展战略，大数据在中国将迎来高速发展期。基于数据思维、数据驱动的理念和实践将是国家、政府、企业、行业制定战略、转型升级、保持竞争力和创新发展的原动力。我们基于在医疗、教育、能源、交通、政务等领域多年的大数据实践，总结出了一些大数据驱动创新的基本理论和方法论，以下逐一阐述。

6.2 大数据创新理论

6.2.1 大数据的宏观性和微观性

大数据具备很多维的特性，而这些特性是数据驱动创新的根本原因。首先，大数据既具备宏观性(也称望远镜特性)又有微观性(也称显微镜特性)，这是清华大学数据科学研究院执行副院长韩亦舜总结的一个心得和体会。宏观性指的是大数据收集的是全样本的历史数据，基于这些数据，可以预测未来。用大数据进行预测，可以预测今后的经济走势、市场形势、发展方向等。Google 的 FluTrend——流感趋势预测，可以基于人们在 Google 搜索引擎

上的搜索关键词，将与流感相关的关键词关联起来，比美国国家卫生署提前一周到半个月准确预测流感疫情的爆发。美国罗切斯特大学的学者和微软公司的研究者一起合作，分析了从703个人和396辆车上收集的超过32 000天的GPS数据，他们从数据中寻找模式并计算一个人某个时间会在某个地方的概率。根据他们的模型，能够预测一个人在未来80周的行踪，并且预测的准确率达了80%。也就是说，根据这些人的历史出行记录，我们可以预测他们未来1年半中的所在的位置，这就是大数据的望远镜特性。大数据的微观性指的是通过精确掌握企业或是个人的最细微的细节，我们就可以通过大数据来做精准画像和服务。美国的电商网站亚马逊，用户80%的再次购买行为都是基于系统的推荐，这是因为系统记录了用户的基本信息以及他们每次的消费信息，包括家庭购物信息，这样就可以准确地掌握他们的行为、兴趣、意图和爱好，从而推测他们会喜欢什么样的商品，为他们提供精准的推荐。大数据就像显微镜一样，观察到了用户最细致的信息，了解他们的一举一动。当然，这里面也涉及到用户的隐私，因此在精准服务和隐私保护两方面要做好平衡。

6.2.2 大数据的生产要素性

大数据的另一个特性是它的生产要素性。大数据之所以能起到革命性和颠覆性的作用，最根本原因就是大数据成为一种新型生产要素。我们以前学资本论的时候知道生产要素有劳动力、资本以及土地等自然资源。传统的生产方式是人加工自然资源，把它们变成产品进行销售，在其中产生增值。当数据成为一种生产要素，加入生产过程时，可以完全替代其他原有生产要素，或是改变原有要素的构成比例。一个简单的例子就是Google的自动驾驶，通过学习和掌握人类的驾驶行为，使用传感器和基于人工智能的自动驾驶软件，可以完全替代最有经验的司机，在这里不再需要司机这一要素了，这就颠覆了出租车行业和驾驶行业。再比如阿里做的阿里小贷，在缺乏数据的情况下，一个传统的银行要放贷的话，需要对贷款的企业进行线下调查，比如说经营状况、员工数量、固定资产、有没有资产抵押等，再进行各种各样的分析，差不多需要一个多月才能放一笔贷款，即使这样也不能保障这个企业可以顺利还款。而阿里通过淘宝、天猫所有平台上面的数据知道商户所有的业务、资金周转、信用等情况，放贷只需要几分钟甚至是更短的时间，在放贷成本和周期上面大大地节约，这就是数据成为生产要素，不需要那么长的时间，那么多的劳动力和调研、金钱来决定是不是放贷。这样的话，传统银行很难和这种新兴的基于数据作征信和风控的新型互联网银行竞争，面临被淘汰出局的危险。大数据作为新的生产要素，正在改变全行业的格局。

6.2.3 大数据的基因特性

大数据的再一个特性，是基因特性。我们知道植物的种子，可以生根、发芽、开花、结果，由一颗葵花子，可以长出一大盘向阳花和无数颗葵花子。人类的胚胎可以孕育出小宝宝，长出头发、眼睛、指甲等不同的身体部位，既像爸爸又像妈妈。同样，一个国家和一个企业的数据，本身是承载着这个国家和企业的基因，这是由基因的遗传性决定的。一个企业通过数据把它整个企业的基因传承下去，但一个企业要根本性地改变它的基因是很难的，这也是很多企业想拥抱互联网，实现转型升级，但举步维艰的原因。基因有一个特性，就是可以进行物种的交叉。如果多种数据源交叉，就好像人种的交叉可以生出混血儿一样，特别聪明，特别

漂亮，多种数据一交叉一融合就可以诞生新的数据，形成新的数据元素，产生很大的一个变革。因此企业如果跨界融合，就可以形成突破创新。还有就是基因有突变。基因如果突变是朝着好方向发展，会得到更优秀的物种和人类。如果朝着坏方向发展就会得癌症。数据如果利用不好就很可能带来毁灭性打击和影响人身安全，利用得好就会诞生全新的商业模式和全新的数据使用方法。

6.2.4 大数据的催化剂特性

大数据蕴藏着巨大的价值，越来越多的人正意识到这一点。大数据被誉为新时代的黄金和石油，然而大数据有一个特性，是黄金和石油不能比拟的，使得它比黄金和石油都更有价值。这就是数据的催化剂特性。我们在初中化学中学到，催化剂可以加速化学反应的过程，但它本身并不损耗。同样，数据在使用过程中可以加速整个生产、经营和商业营销的过程，但数据本身并不损耗，怎么用数据都是在那里。数据可以重复使用，而且数据还可能越用价值越高。数据跟多种数据源交叉使用的时候价值沉淀就越来越大。大数据可以深入到全行业，可以循环使用。任何一个行业要素都会损耗，用完就没有了，但数据可以一直用它，越用越值钱，可以说是最值钱的行业，这是其他生产资料不可比拟的。

6.2.5 大数据的活性和流动性

前面提到了数据的很多好的特性，但是如果光有数据，不把它们很好地利用起来，数据的价值就得不到发挥，所以还要关注数据的活性和流动性。现在社会和企业的数据，已经非常庞大，尤其是一些传统企业，比如医疗行业、银行、交通，累积了几十年的数据，但这些数据，有的是纸质的放在文件柜里，有的是放在计算机中只是用来形成报表，做最基础的统计分析，数据没有利用起来，我们说它们处在沉睡的状态。数据需要活动起来，唤醒起来，才能发挥其巨大的威力。数据也和资金一样，需要周转起来，发挥其流动性。做生意的都知道，资金周转越快，周转的次数越多，就越能赚钱。数据也是一样，需要更快地更多次数地使用数据，才能更多地发挥它的价值。

6.2.6 大数据的黑洞效应和核聚变效应

依据以上的大数据的几重特性，我们总结出大数据具备两个效应。第一个是大数据的黑洞效应，我们知道一个大质量的星体不停地旋转，就能形成强大的吸附力，把周边的物质都吸收进去，甚至连光线都不能逃逸，最终形成一个黑洞。如果整合多行业多源的数据，发挥其活性和流动性，数据的质量越来越大，数据流转速度越来越快，就可以把周边所有相关的数据、资源、人才等都全部吸附过去，形成一个巨大的数据黑洞，最终只要跟这个数据黑洞发生交集的都会被吞噬进去。我们预测未来全球就像宇宙一样，可以形成多个数据黑洞，现在的BAT(百度、阿里、腾讯)，由于其本身累积了大量的数据，同时又在不停地整合行业数据，已经具备了成为数据黑洞的一些条件。大数据另一个效应就是核聚变效应，当多种数据源进行聚合的时候可以产生密度更大、质量更大的数据粒子，这个聚合的过程就是一个核聚变过程，最后能释放出来巨大的能量。全行业的全国性的、全球性的数据聚合起来可以爆发核能量。所以说大数据是新时代创新的原动力、核引擎。

美国的政府数据开放网站 www.data.gov 一年带动的创新产值是 3 万亿美金，这就是

数据聚合所能产生的能量的体现。其中一个例子是美国的 Climate 公司，基于上述政府开放数据网站，汇总了 250 万个地点的气象测量数据和各个主要气候模型的天气预报，同时综合 1500 亿个土壤观测记录，对这些数据进行处理，生成出 10 万亿个天气模拟数据点，为农业生产提供保险服务。Climate 几位联合创始人是谷歌的早期员工，他们为天气保险的投保人开发了一种自助式服务，此前这类保险只能通过定制的方式进行柜台交易。现在，客户可以登录 Climate 公司的网站，确定特定时间段内需要投保的气温和/或降雨量范围。平台收到订单后，就会在 100ms 内综合分析天气预报、近 30 年来的国家气象局（National Weather Service）数据，以及用户所在地的地质调查数据，并根据气候变化，对分析结果进行微调。得出结果后，就会作为保险商，给用户开出保费。投保人如果因为意外天气而受到损失，就能自动获得赔偿。Climate 公司最终被美国最大的农业公司孟山都以近 10 亿美金收购。

6.3 大数据创新方法论

综合以上的大数据的特性和效应，相信很多人都已经认识到大数据的威力和前景，希望在大数据行业进行创新创业。但是具体选择什么样的行业，采用什么样的商业模式，如何判断创新创业是否能成功，是大家共有的问题。我们依据大数据的行业实践，总结了一些大数据的创业方法论。我们制定了 10 个维度和指标，来指导和衡量创业创新的方向和方法。这其中有 5 个基础指标，我们认为这些指标缺一不可，是成功的基础。另 5 个是重要指标，是指这些指标很重要，可以让数据创新以爆炸式的发展模式进行扩张。但是也允许局部指标缺失，需要尽力去考虑和满足。

1. 基础指标

（1）价值密度：产业链上单位时间内创造的产值。产值越大，密度就越高，数据创新所承担的风险度也就越高，成功的可能性就相对较大。

（2）基础约束度：体制、机制、政策、资本等约束，常常是能否实施的关键。约束越大的场景，数据的流动性就越受约束，作用就越小，所带来的创新和变异就越少，信息化推广和建设的阻力就越大。比如医疗领域，行业门槛和阻力就较大，不容易形成数据创新和突破。

（3）投资收益度及公益度：投资收益好，积极性就高。或者公益性好，政府扶持力度就大。二者如能结合，则最佳。数据创新是逐利性和公益性并存的，这个利不仅表现在经济利益上，也表现在社会认可上。

（4）市场接受数据的粒度与敏感度：粒度指的就是粗糙度，粗糙就是颗粒度大，精细就是颗粒度小。很粗糙的数据也有人买单，就意味着对数据的敏感度弱。如中国的教育，只要说出对孩子有帮助，多差的产品，也会有大量的消费人群。

（5）数据的全量度与实现应用的速度：互联网时代比拼的是谁更快，谁覆盖的更广，更有执行力和实现力。但实现的速度和数据采集的全量度会产生矛盾，关注了数据的全量度，就势必影响实现的速度，反之亦然。怎么样实现这两个指标的协调统筹，往往成为成败的关键。

2. 重要指标

（1）用户群与地理区域覆盖度：市场和数据覆盖得越广，周旋空间和转型路径就越多，企业和产品就越安全。

（2）行业技术门槛高度：技术与人力资源的要求，不是基本约束条件，但如果形成技术门槛，追随者一时难以赶上；没有突破，就很可能被取代。

（3）社会经济发展支撑度：经济发展的程度，往往代表人们信息消费的力度，但可以采取适当的跨越式发展。经济越发展，数据创新越容易被接受。

（4）行业关联、渗透与应用维度：行业内应用维度的多少和行业外渗透关联力度的大小，往往是爆发式增长的前提和保障。

（5）原有行业规模与竞争激烈程度：竞争越激烈，模式和内容的创新需求就越强烈，切入的机会就越大，形成的效果就越显著。

那么如何运用这些指标呢？简单来说，首先就是选择数据价值密度高的行业去创业。上一个时代价值密度高的是房地产行业，现在按照我们的分析是金融行业、健康行业、教育、旅游这些行业，价值密度很高，每年每个用户花费上万元。但是行业价值密度高不一定做得顺利。第二条要看行业门槛够不够高？可不可以进去？比如要进入医疗行业，把所有数据都打通，把全国所有医院的癌症片子都拿过来汇总，将产生巨大的价值。但是，医院不可能随便把数据拿给你。所以，在医疗行业突破这个门槛就非常困难。其他的维度也可以照此分析。

上面定义了5个基础维度和5个重要维度，这些维度的衡量，可以用类似于图6-2的蜘蛛网状的重心图，如果每个指标都是相对比较均衡，分值较高，那么创新创业的成功几率就比较大，如果在某些维度严重缺失，那就要考虑调整方向，或是如何弥补相关的缺失，否则创新创业的道路就会比较艰苦。本书作者希望通过这样一个简单的方法论指导读者在大数据创新创业方面去做尝试。

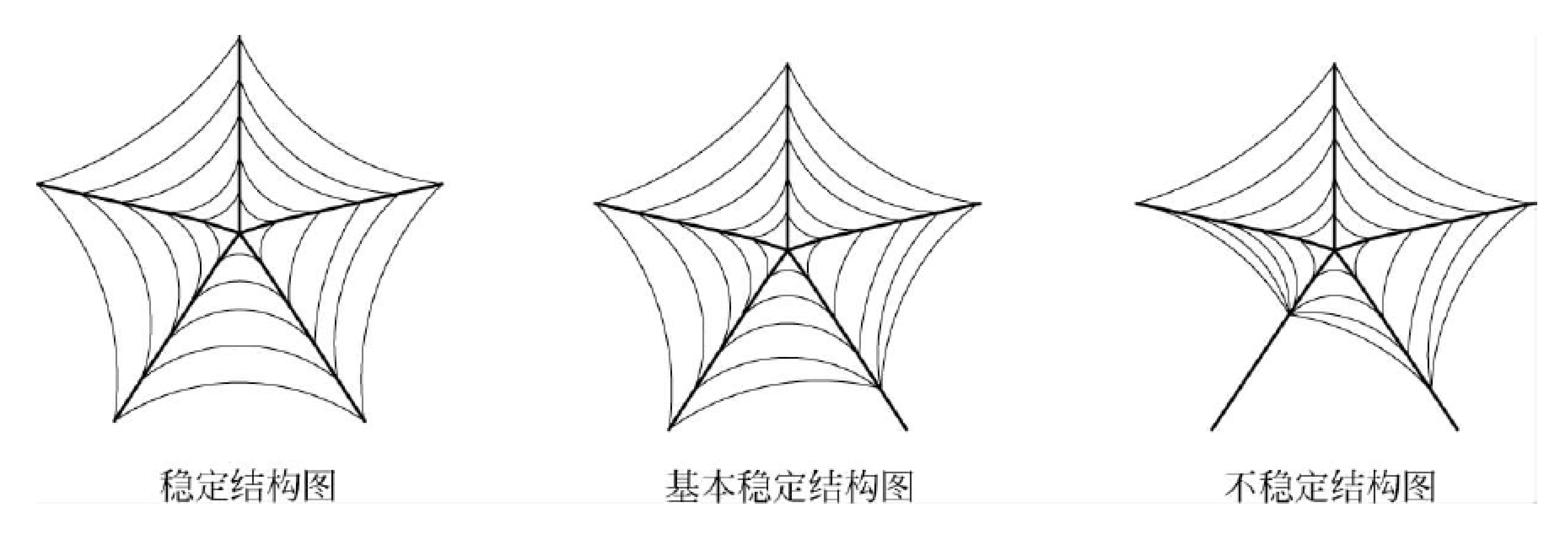


图6-2 稳定性结构蜘蛛网图示例

6.4 信息演变趋势

从总体的信息科技和互联网的发展趋势，我们总结有3个阶段。现在我们国家提倡“互联网＋”，把“互联网＋”模式应用到所有的传统行业里面去。从本质上来说，“互联网＋”改变的只是边界和渠道，也就是说，原来没有进行传统线下行业经营的，可以通过互联网的方式来介入这些行业，并且改变行业生态。另一方面，传统的市场、经销、营销渠道，转移到了基

于互联网的渠道,需要通过新一代互联网媒体、社会化媒体来到达受众。但是,“互联网＋”并没有改变商业的本质,互联网应用最终还是要应用在数据上面,应用在企业商业模式上,依托数据形成革新,才能形成根本性的变革。所以“互联网＋”发展的下一阶段必然是“大数据＋”,将大数据和各行业结合起来,深度利用数据,发掘数据价值,才能形成理论、模式、技术和应用方面的创新。那么“大数据＋”发展的下一步是什么呢？数据分析挖掘的目的是什么？目的还是要产生类同于我们人类的智慧。所以,再往后面发展就是所谓的机器智能和人类智慧相结合的这样一个时代。现在英国科学家已经把人脑和计算机合在一起,连通了,也就是我们所说的“奇点”时代正在来临。未来的决策我们会分不清楚,可能有一半是自己想的,另外一半是计算机做出来的。机器智能最终能做到替代高级劳动力的作用,我们已经在自动驾驶中将司机替代了,未来会替代高级医生、教师、律师、投资顾问等,而且这些都正在发生。IBM 的 Watson 机器人已经应用在癌症诊断方面,因为处理的案例和数据量大,比最有经验的医生诊断率还高百分之十几。

6.5　大数据创新实践闭环

大数据在行业创新和实践,不是简单的数据采集和分析,尤其是站在国家发展战略和企业决策的层面上,需要形成理论、创新和实践体系。那么如何在各个行业领域进行大数据创新实践,如何让大数据在行业里具体落地和发挥价值？下面介绍一个简单的创新实践闭环。

1. 系统科学的理论和方法论指导

大数据的研究和应用离不开科学理论的指导。首先需要基于数据科学和数据创新理论指导。数据科学横跨多个学科领域,要形成对数据的洞察、数据源及属性的选取、数据模型的选择、所采用的分析及验证方法,都需要系统、科学的理论指导和方法论。基于前述的大数据创新理论和方法论,可以有效地进行大数据的行业融合和模式选择。

2. 标准体系的建立

大数据处理的最多的就是多源多格式的数据关联分析,在理论指导的基础上,对于数据的表示、存储、处理、交换、共享、展现等都需要建立标准体系。只有建立在开源、开放的平台上,有相应的数据访问标准及接口,才能真正促进数据的互连互通,发挥大数据的威力。目前在云计算和大数据领域,相关的标准建立都还在起步阶段。美国的国家标准与技术研究院 NIST 成立了一个大数据工作组,致力于大数据的标准制定。同时由欧盟委员会、美国政府及澳大利亚政府发起组织的研究数据联盟(Research Data Alliance)也在进行数据方面的标准制定。中国计算机学会的大数据专家委员会也是致力于大数据标准制定的专业组织。

3. 合理的人才和知识储备

数据科学的研究及应用都离不开数据科学家和数据相关的从业人员的参与和贡献。数据科学家是 21 世纪全球抢手和紧缺的人才,因此人才、知识的储备、教育、培养和培训就尤为重要,掌握了人才和知识才能在“数据为王”的新时代占领制胜高点。美国政府推出的面向高级工程专业的移民政策吸引了世界上一大批优秀人才,各大高校也在纷纷开设数据科学专业及课程。我国也应加强数据科学相关专业人才的政策吸引,打造创业环境;在高校课程及专业设置及建设方面急需加强;企业更是要创造良好的人才及培训环境,注重全员

的大数据培训，才能在大数据浪潮中不被淘汰。随着我国大数据战略的实施，很多地方政府都制定了大数据人才的培训培养战略，各类高校、职业学院/学校、培训机构也都开设了大数据人才的培训和实践课程。大数据人才的储备是大数据产业发展的基础。

4. 典型应用场景的分析

由于大数据是在现实生产场景中遇到的切实问题，因此大数据的应用不能走主观、脱离实际的道路。要到生产一线中去发现问题，分析实际应用场景中已经不能解决或是急需解决的大数据问题。正是由于实践问题，才能驱动大数据的技术应用和技术创新。例如城市的交通视频监控，一个中等规模的城市每天产生的视频数据就达十几 TB，在数据的存储以及实时分析方面就面临巨大的问题，将大数据应用于这些领域，就能马上产生价值。同时注重从数据出发，梳理数据资产和发挥数据的融合效应，将有效地形成应用创新。

5. 核心关键技术的研究

当前得到广泛应用的大数据技术还是以 Hadoop 为主的开源技术，开源技术在大数据生态中将占主导地位，也对行业做出巨大贡献。但基于开源技术带来的挑战，是技术门槛降低和激烈的竞争，当前 IBM、微软、Intel、Oracle、HP 等 IT 巨头都推出了基于 Hadoop 的大数据集成产品。因此在大数据行业中，还是要结合行业知识、经验和实践，形成企业自己的核心关键技术，同时加强技术运营、维护及服务，才能提高企业的竞争力，在大数据市场中占领一席之地。

6. 自主可控的产品

中国的信息化建设，长期处于被国外先进产品和技术垄断的状态。大多数政府及企业的信息化架构，都是基于 IOE 三驾马车，即 IBM 的服务器、Oracle 的数据库，以及 EMC 的存储。采用国外成熟先进的产品，本也无可厚非，但确实对国产自主的软件开发及行业发展造成了极大的阻碍。随着大数据的爆发，这些大公司的产品本身对大数据的处理能力也都存在很大的局限，加上在国产化和国家信息安全方面的注重，目前全国“去 IOE 化”的呼声越来越高。在大数据时代，目前还未形成占据市场垄断地位的大数据巨头公司和产品，同时开源技术也很普及，因此抓住时代机遇，加强我国自主的关键技术研究，形成自主可控的大数据产品，将使我们有机会实现弯道追赶，打破国外技术垄断，发展我国的大数据产业，在国际市场中一较高低。

7. 开放的创新体制

大数据需要开源、开放的数据、标准和平台，形成开放的实践和创新体制。在此基础上可以集众人之智，采众人所长，形成新技术、新产品、新模式、新服务，促进科技创新和发展。也只有基于开放的体系，才能鼓励和推动创新，促进大数据产业的良性发展。基于数据建设全国性的、地方性的，以及行业和企业级的数据开放、众创、交换和交易平台，将极大地促进数据创新，发挥数据的社会和经济效益。

6.6 中国创新创业大数据版图

为了将前面总结的理论及方法论付诸实践并进行检验，清数科技在成都还建立了全链条的创新创业孵化器——第五维国际大数据孵化器。孵化器为创新创业企业不只是提供办

公场地和创业辅导，还提供云计算和大数据技术平台支撑，依托大数据相关协会和联盟进行市场推广，以及线上线下媒体进行市场营销，同时还设立大数据产业基金提供投资服务。孵化器也建立了和美国硅谷及西雅图、英国的孵化器合作，致力于打造国际级的专业大数据孵化器，帮助创新创业企业成长。在建设和运营孵化器的过程中，我们认识到如果能够利用大数据对全国的创新创业态势进行综合的分析和展示，将不仅帮助我们自己对全国和各省市的双创发展有一个全面的把握和比较，还能够帮助政府、企业、行业和其他各参与方也有同样的收获，因而我们规划并开发了中国创新创业大数据版图，将中国自 2013 年以来的双创产业发展全面、综合地用大数据的方法收集、处理并展示出来。

6.6.1　大数据时代的数据管理

随着云计算、物联网等技术的兴起，数据正以前所未有的速度在不断地增长和累积，大数据时代已然来到。在大数据时代，数据仍然是最关键的。如何将大数据管理好，仍然是对企业的考验。手机通话、移动在产生数据，ATM 在产生数据，商品上的 RFID 在产生数据，包裹从一个城市到另一个城市在产生数据。就算是一个小小的店铺，当它销售出去一瓶水，也可能会记录到 Excel 里面，产生数据。数据记录着世界的存在和变化。

当企业的某项资产非常重要，数量巨大时，就需要有效管理。如今，数据已经成为这种资产。以前人们还不会将它看作是资产，而是一种附属物。客户来办理业务，在系统中产生了这种附属物。而现在，发现在客户办理业务这条信息中，蕴含着一些客户的需求，成千上万条这类信息累积下来，就能洞察客户所需，为设计新产品，为客户个性化营销产生新的价值。数据变成了一种资产，需要被管理起来。

人类历史上从未有哪个时代和今天一样产生如此海量的数据。数据的产生已经完全不受时间、地点的限制。从开始采用数据库作为数据管理的主要方式开始，人类社会的数据产生方式大致经历了以下 3 个阶段。

(1) 运营式系统阶段。人类社会数据量第一次大的飞跃正是建立在运营式系统开始广泛使用数据库的基础上。

(2) 用户原创内容阶段。互联网的诞生促使人类社会数据量出现第 2 次大的飞跃。

(3) 感知式系统阶段。人类社会数据量第 3 次大的飞跃在于感知式系统的广泛使用。随着技术的发展，人们已经有能力制造极其微小的带有处理功能的传感器，并开始将这些设备广泛地布置于社会的各个角落。这些设备会源源不断地产生新数据。

数据的产生渠道变得更加广泛，同时数据对于政府和企业的重要性愈来愈强，如何收集和管理这些数据就成了人们广泛关注和研究的问题。大数据特有的 4V 特性让以传统关系型数据库作为核心的数据管理方式变得不再有效，面对海量异构的数据时，关系型数据库显得越来越力不从心，仍然以第一阶段的运营式系统的方式来建设大数据系统是不现实的。而政府和企业对技术的需求变得更加强烈，所以各类大数据技术开始迅速发展。从数据采集、数据传输、数据存储、数据分析到数据可视化等各个环节都有新的技术和框架不断推出，去适应和解决大数据环境中的各种问题。

6.6.2　大众创业万众创新的浪潮

最早在 2014 年 9 月的夏季达沃斯论坛上，李克强总理在公开场合发布“大众创新、万众

创业”的号召。他提出要在960万平方公里土地上掀起“大众创业”“草根创业”的新浪潮，形成“万众创新”“人人创新”的新态势。此后他在首届世界互联网大会、国务院常务会议和各种场合中频频阐释这一关键词。每到一地考察，他几乎都要与当地年轻的“创客”会面。他希望激发民族的创业精神和创新基因。

2015年，李克强总理在政府工作报告中又提出：“大众创新，万众创业”。政府工作报告中如此表述：推动大众创业、万众创新，“既可以扩大就业、增加居民收入，又有利于促进社会纵向流动和公平正义。”在论及创业创新文化时，强调“让人们在创造财富的过程中，更好地实现精神追求和自身价值”。

在信息经济发展当下，知识经济、共享经济、创新经济成为时代潮流，尤其当90后逐步成为消费市场的主体。无论是欧美等发达经济体，还是中国，“互联网＋”的发展势头都锐不可当。“互联网＋”是衡量创新创业的合理的切入点，“互联网＋”样本数据既能有效识别创新驱动，又能有效地跟踪创业者的创业不同阶段。更为重要的是，在中国，它代表了产业发展的未来趋势与走向。而“互联网＋”和双创的有机结合，为中国经济转型和改革提供了源源不断的创新动力和创造活力。对中国主动适应和引领经济发展新常态，形成经济发展新动能，实现中国经济提质增效升级具有重要的意义。

中国的经济发展经过30年高速发展，在“十三五”期间开始进入新的发展阶段，社会发展对创新的要求提高到了一个更高的层次。根据工商总局公布的数据，2016年第一季度全国社会投资创业势头良好，新产业、新业态、新模式蓬勃发展，特别是小微企业活跃度稳步提升。

当前国内的创新创业形势一片大好，特别是各级政府都出台了大量政策吸引和鼓励创新型企业的入驻及发展。中央各部委更是多次发出指导性文件，为双创服务的政策体系正在逐步完善中。有数据显示，目前，全国各类众创空间已超过2300家，与现有2500多家科技企业孵化器、加速器，11个国家自主创新示范区和146个国家高新区，共同形成完整的创业服务链条和良好的创新生态，这些众创空间、孵化器、高新区共同构成了为双创企业服务的完整链条，为双创企业的发展提供坚固的平台。

但是在大好形势的背后也出现了一些问题，2015年下半年开始，双创的这股热情似乎有些低落，产业界中频频出现各种创业企业的死亡名单，“投资的冬天开始出现”等论调也开始出现。各种资本、孵化器等产业扶持力量的介入，似乎并未带来预料的理想效果。当前，创新创业项目出现低潮，原因是多方面的，其背后存在着团队、产品、资金、市场、产业等诸多问题，主要体现为创业团队的清晰定位、对待投资的客观态度、产业环境的准备把握三大方面。

当梳理出创业失败背后的原因之后，创业团队要审视和思考自己的优势在哪里，如何发挥出这种优势？创业过程中哪些环节可能会出现问题，如何规避和化解这些问题？又该如何学会借力？资本、产业资源又该如何发力，才能有效帮助到创业团队？政府要如何灵活地修订政策，才能释放创业团队的活力？这些问题都是亟待解决的重要问题。

6.6.3 中国创新创业大数据版图的推出

那么，应该如何去解决这些问题，发现真正的创新创业价值，释放大众的创业激情和创

新能力呢？我们需要从数据的角度去全景式地掌握全国各个地区的双创发展状况，衡量当前火热的双创态势下企业的真实生存状态，并为创业企业提供深入有内涵的市场分析。这样才能准确地发现创业的痛点，做到资源的合理分配，政策的合理制定，人才、资金的合理流动，这就涉及如何利用双创数据的问题，而这正是一个典型的大数据问题。

全国双创相关数据包括工商数据、政府数据、市场数据、媒体数据等各种数据来源，而且数据的类型和产生速度都不一样。如何解决各种数据源的融合，将双创相关数据利用大数据技术来进行有效的管理利用，是需要主要解决的问题。

得益于大数据的全面性、完整性，我们可以同时以宏观和微观的视角去审视当前全国的创新创业形势，发现潜伏其中的问题和机遇。以往我们都是通过政府报告或者新闻媒体报道的形式去了解当前的形势，但是总会存在宏观数据无法深入传递数据价值，宏观数据掩盖细分领域发展情况的问题。而微观数据则面临无法让人总揽全局，容易陷入特定案例情况，或问题定位错误的情况。

清数中国创新创业大数据版图是利用目前领先的大数据技术，基于清数自主研发的大数据一体机 NEO，结合深度的调研摸底，对全国海量双创数据多维度的采集、储存、分析、挖掘、可视化的全流程处理，展现了一个实时更新、覆盖面广、参考价值高的全景式版图，主要特点就是多维度、可比较、相关性强、全景式，是大数据技术在应用层面的一个集中展示。

目前清数双创大数据版图已经涵盖了双创核心二十多个维度，收录全国所有地区从 2010 年至今的双创相关数据，并进行集中、全面的数据深入分析。同时对不同维度相关性数据融合后，利用算法，分析得出清数双创指数来整体反映一个地区的双创活跃度（如图 6-3 所示），同时还能进行省份、城市及地区之间的双创指数对比，为政府、企业、投资、创业、就业决策提供综合的参考依据。

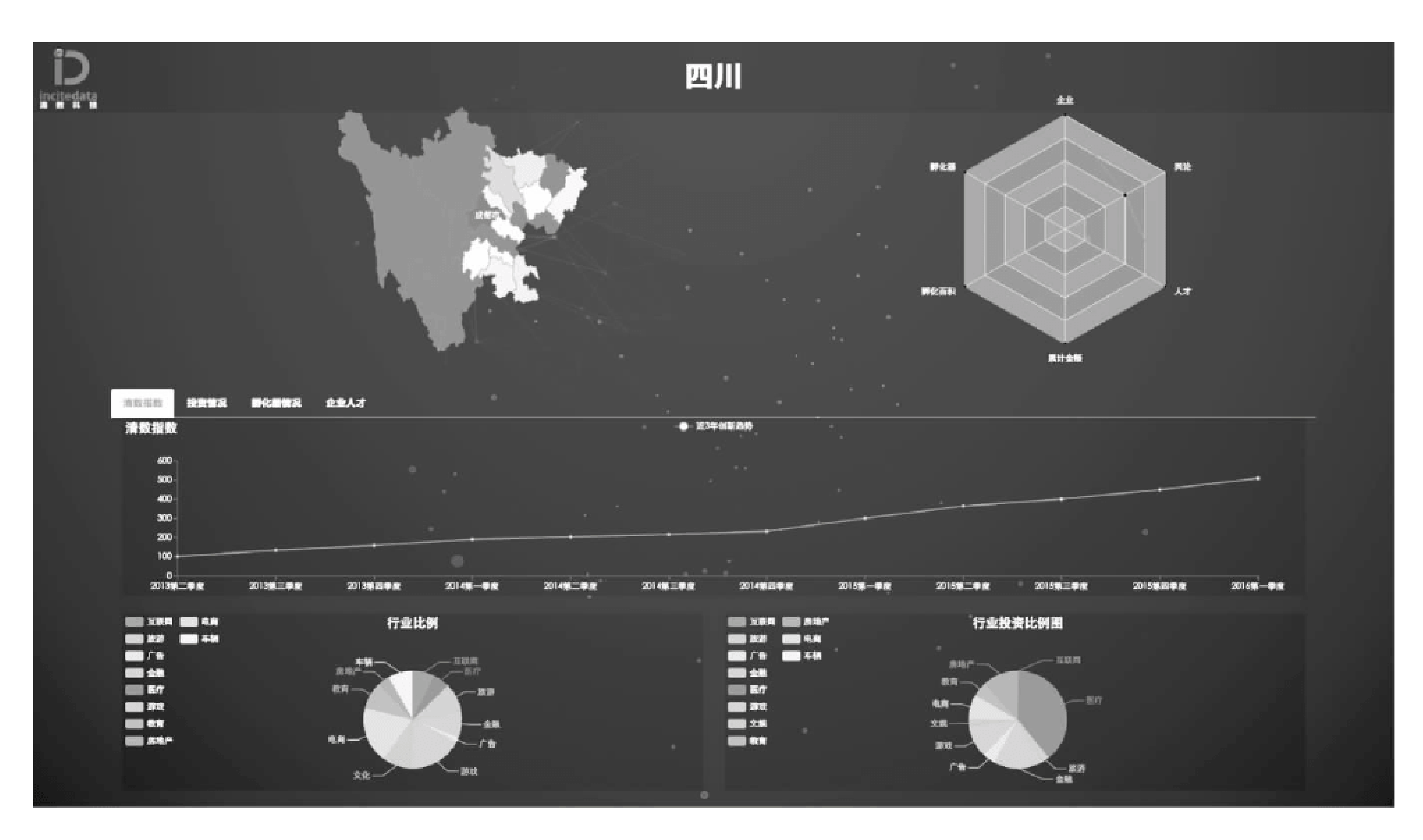


图 6-3　中国创新创业大数据版图——省份统计

目前架构中对于双创关注的核心维度已经全部上线，发布之后随着数据的补充更新和局部细化，最后能将版图细化到某地市某区的某个园区和企业，达到对双创数据的全方位管理，真正实现双创数据价值的转化，目标是把双创大数据版图做成一个关注和致力于双创产业的基础检索和辅助决策工具。

6.6.4 双创版图中的大数据管理挑战

针对双创数据的数据管理面临很多挑战需要解决，在中国创新创业大数据版图的实现过程中，综合利用多项大数据技术逐项突破，实现了全国双创数据的集中管理和利用。

那么主要面临的各项挑战又有哪些呢？我们总结有以下几点。

1. 数据集成

数据的广泛存在性使得数据越来越多地散布于不同的数据管理系统中，为了便于进行数据分析需要进行数据的集成。数据集成看起来并不是一个新的问题，但是大数据时代的数据集成却有了新的需求，因此也面临着新的挑战。大数据的多源异构特性决定了要整合各类数据源，并处理好数据爆炸问题。

2. 数据质量

数据量大不一定就代表信息量或者数据价值的增大，相反，很多时候意味着信息垃圾的泛滥。一方面，很难有单个系统能够容纳下从不同数据源集成的海量数据；另一方面，如果在集成的过程中仅仅简单地将所有数据聚集在一起而不做任何数据清洗，会使得过多的无用数据干扰后续的数据分析过程。大数据时代的数据清洗过程必须更加谨慎，因为相对细微的有用信息混杂在庞大的数据量中。如果信息清洗的粒度过细，很容易将有用的信息过滤掉。清洗粒度过粗又无法达到真正的清洗效果，因此在质与量之间需要进行仔细的考量和权衡。

3. 数据处理的实时性

随着时间的流逝，数据中所蕴含的知识价值往往也在衰减，因此很多领域对于数据的实时处理有需求。随着大数据时代的到来，更多应用场景的数据分析从离线转向了在线，开始出现实时处理的需求，比如实时广告竞价问题。大数据时代的数据实时处理面临着一些新的挑战，主要体现在数据处理模式的选择及改进。在实时处理的模式选择中主要有 3 种思路：即流处理模式、批处理模式以及二者的融合。各种工具实现实时处理的方法不一，实际应用中往往需要根据自己的业务需求和应用场景对现有的这些技术和工具进行改造才能满足要求。

4. 隐私数据的保护

很多时候人们有意识地将自己的行为隐藏起来，试图达到隐私保护的目的。但是互联网尤其是社交网络的出现，使得人们在不同的地点产生越来越多的数据足迹。这种数据具有累积性和关联性，单个地点的信息可能不会暴露用户的隐私，但是如果有办法将某个人的很多行为从不同的独立地点聚集在一起时，他的隐私就很可能会暴露，因为有关他的信息已经足够多，这种隐性的数据暴露往往是个人无法预知和控制的。从技术层面来说，可以通过

数据抽取和集成来实现用户隐私的获取。而在现实中通过所谓的众包方式往往能更快速、准确地得到结果。

6.6.5　双创版图中大数据技术的集中运用

首先需要对双创代表的含义进行建模，确定什么样的维度能够反映出双创的形势。比如要了解某个地区的人才情况，需要有不同层次人才的详细人数，更深入的可以有留学归国创业人员的留学国数据，以此来判断哪个国家有更多的留学生回国，及哪个国家的留学生更有创业精神，如图 6-4 所示。

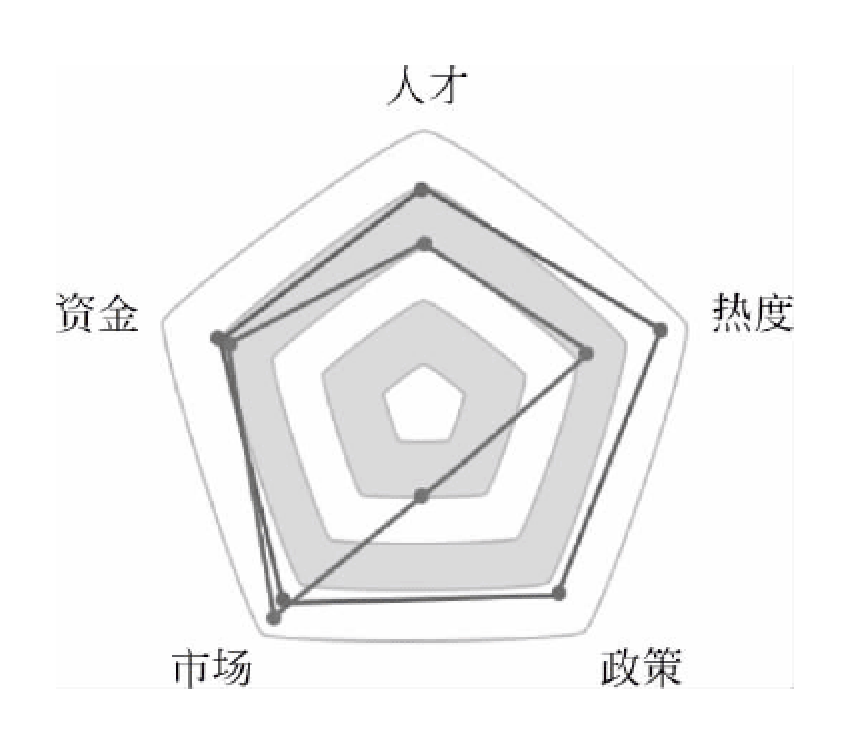


图 6-4　清数双创大数据指数

为了能够定义出更精确、更广泛的维度，我们进行了大量的讨论和对实际情况的调研。最终将双创相关的维度划分为 5 个大方向：资金，人才，政策，市场，热度。资金代表了市场和政府资金的数量和走向，是对创业起到决定性作用的一个要素，哪里有资金哪里就有活跃的创业企业和创业团队，资金是最代表资源配置情况的一个维度方向。人才则代表了当前人才的聚集和流动情况，一定程度上人才是和资金的情况具有正向关系的。政策是代表了政府对于双创的政策支持和政策执行情况。市场表示的是当前双创企业相关的市场指标，比如总体企业数量、当月新注册企业数量、分行业企业数量等。热度则是反映双创在新闻媒体和网友群众之间的讨论度及关注度，有更多人讨论的企业或者产品表示了其目前正处在迅速扩张的阶段，也更容易获得资金和人才的青睐，同时可以结合政策和市场这两个维度，看出企业获得的支持力度如何，是否正在引领一次新的细分行业创业浪潮。

在定义出了相关的能够反映双创的维度后，需要获取的是对应的数据。当前的这些维度，需要获得的数据分为两类，一类是可以通过公开渠道获取到的，比如新闻媒体的报道和网友的讨论，另一类是需要通过政府等相关数据源的合作建设才能获取到，这一部分我们采取了循序渐进的方式来建设，通过和多地的政府建立合作关系来逐步丰富，最终获取到接近全量的数据。如图 6-5 所示为双创版图的基本数据处理流程，可以看到我们对通过各种数据收集方式获得的数据经过一个完整的数据分析挖掘流程以后才能得到最终的数据产品来呈现给终端用户。这其中的每一步都是对数据的过滤和价值提炼，让数据能够融合，产生更多的聚集效应。

数据的收集是一个非常重要的步骤。我们为了获取到尽量多的数据，先采用了在公开网络上抓取的方式对相关的数据进行获取。在公开网络上使用爬虫抓取数据已经是一个很

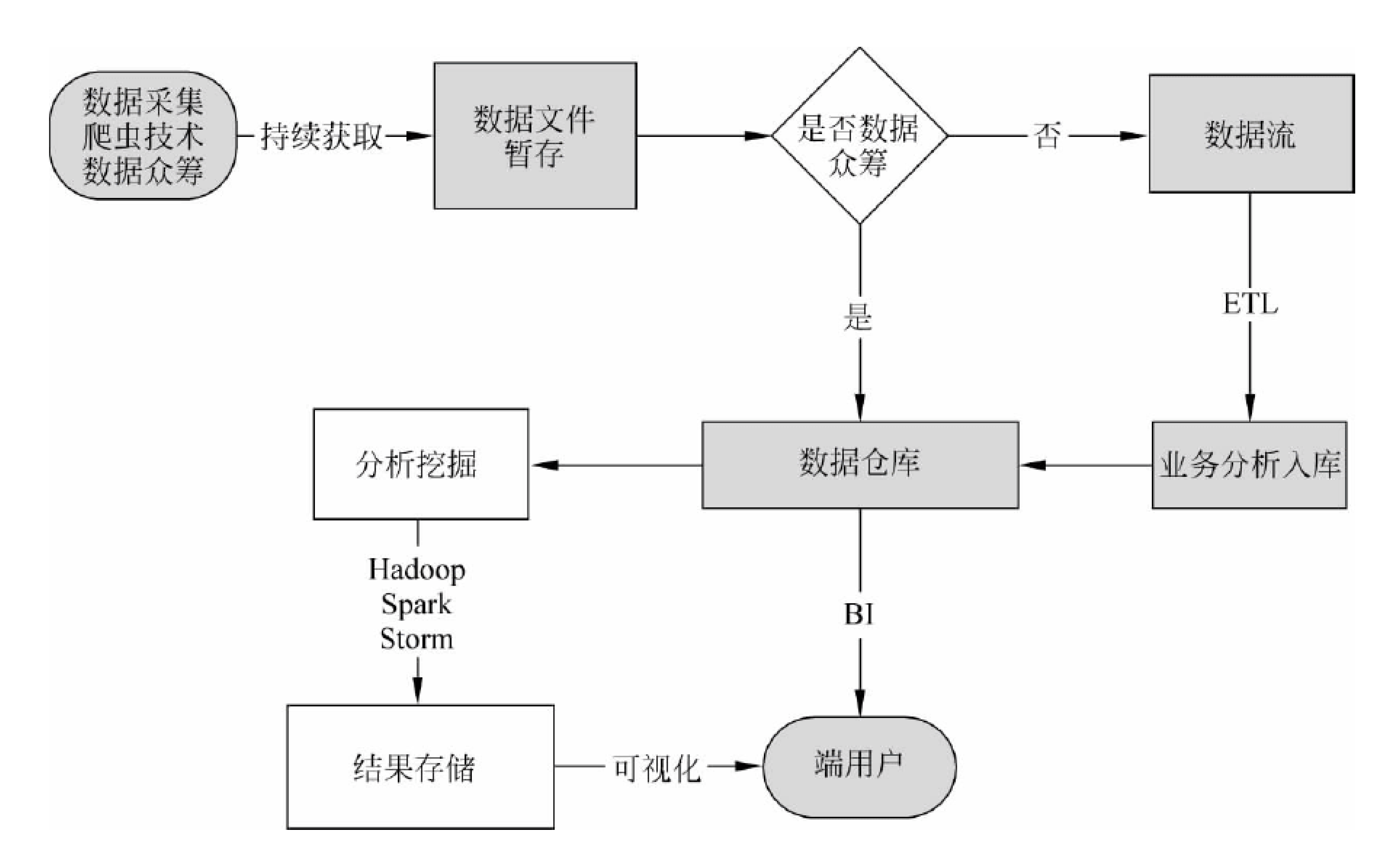


图 6-5 双创版图的数据处理流程

成熟的技术，而且是 Google、百度等搜索引擎的基础技术。成千上万的爬虫没日没夜地在互联网上爬取数据，才造就了这些企业海量的数据基础。这次我们的数据抓取工作主要是针对国内的新闻媒体的报道和微博论坛等公开网络空间言论。

我们自己编写了针对性的爬虫机器人，就像是人在浏览网页一样，不仅对互联网上的指定内容进行浏览，同时将浏览过的网页保存到爬虫机器人所在的本地存储中。这些数据经过规整入库以后，能够形成一个信息和文本检索库，这就是我们分析的一个重要来源。比如说一款新的社交 APP 在微博上的讨论次数、点赞次数、分享次数，就可以直接从本地检索库中获取到，从而根据这些次数来判断出这款 APP 当前的讨论热度和上升趋势。

另一方面，我们也采用数据众筹的方式，和地方政府、科技园等合作，获取当地科技企业的完整数据。这部分具有权威性的数据也是重要的种子数据，可以为指数计算中的权重设置提供重要的依据。

类似地，我们采用同样的方式对其他需要的数据都进行了抓取，规整入库后，就有了分析的基础。一个公司的基本信息、创始人信息、融资信息、新闻媒体和网友对其的讨论和评价等，构成了一个企业的全方位全角度的数据解读，再结合其所在行业和地区的市场和政策情况，就能定义出企业的健康情况、活跃程度、发展爆发力等更有意义的维度。对所有企业都能够进行量化的判断后，就能从行业、地区等角度总结出整体的双创态势，更进一步地可以用指数形式来描述，从而进行企业、地区和行业间的量化比较和评级。

接下来在对双创数据进行处理的过程中，我们使用了自主开发的 NEO 大数据一体机。NEO 开发团队通过总结多年的大数据实施经验，整合行业专家多年的算法经验，总结应用的共性，沉淀多个行业模型，提炼出一整套大数据实施标准，针对大数据中常见的数据分布、数据连接等问题，提出处理准则，有效解决了大数据实施中常见的问题。

NEO 大数据一体机正是这些经验沉淀的产品，区别于 SAP Hana、Oracle RAC、Exadata 等高硬件配置的一体机产品，NEO 通过软硬件联合优化的思路，从网络 I/O、硬盘、

内存使用、数据感知多个方面着手，尽可能发挥软硬件能力，从而极大降低大数据实施成本。NEO 一体机通过标准化的架构集成服务器、存储、网络、软件、操作系统等配置，简化工程实施难度，简化数据中心基础设施部署和运维管理的复杂性，提升服务器运行效率、降低部署调优难度、避免差异化设备导致的集群不稳定性，为行业用户提供成熟的一整套集成解决方案，让用户将研发重心放在用户应用业务的开发上，增加商业附加价值。

在使用 NEO 大数据一体机过程中，可以专注于设计分析模型，而不用再关心集群性能是否发挥完全、算法是否已足够优化等问题，这些都由 NEO 大数据一体机进行集中的处理，其最后提供给分析人员的是统一的分析工具和能力。

我们主要是利用了 NEO 大数据一体机提供的海量存储扩展能力和多维度分析查询。在对公开网络上的数据进行抓取时，其数据量是不可预测的。可以利用 NEO 大数据一体机的无限扩展的特性，从规模较小的集群开始搭建，根据抓取到的数据实时地扩展集群规模，同时不影响数据抓取工作的进行，这极大地提高了工作效率和工作难度。

在对数据进行深入的分析时，我们利用 NEO 大数据一体机的多维度分析查询功能，针对我们目前定义的二十多个维度设计了更多的交叉维度分析和查询功能，NEO 能够很好地支持实时的分析和查询，同时能够在秒级甚至毫秒级完成。在对数据进行各种横向和纵向的分析时，能够实时产生分析结果，实现了分析角度的极大扩展。

最后，通过数据可视化技术将分析结果进行统一的展示。这里选择了更有定制性和扩展性的数据可视化技术，能够根据我们的需求来生成分析结果的呈现效果，并且能够以全景、深度、直观的方式来展示最后的分析结果。

清数双创指数是我们在分析过程中提出的最新概念。类似现在的各种经济指数，双创指数是整体反映一个地区的双创健康程度和活跃程度的综合指数，以量化的方式直观地体现地区双创竞争力。清数双创指数的提出首先是对目前的二十多个维度的数据进行类型、性质、影响程度的划分，然后根据维度的各种不同属性在现有数据基础上进行数据建模，构建出维度计算的框架，最后通过 NEO 大数据一体机提供的实时计算能力，不断地从抓取到的数据中进行计算，从而得到一个不断更新的指数。指数的计算过程包含对所有维度的实时计算，而且需要在秒级进行更新，这对计算集群的性能是一大考验。NEO 大数据一体机很好地完成了这个工作，并基于标准化的实时计算能力提供，极大地简化了数据模型构建时间和工作复杂程度。

经过上述的数据采集和处理过程，并经过可视化呈现，我们最终得到了中国创新创业大数据版图。在经过细致的分析和处理后，数据已经展现出不同的价值，从最初混杂的价值密度较低的数据变为统一的直观的高价值密度数据，同时形成数据采集、处理、分析、呈现的完整实时链条，并且能够提供给使用者进行更进一步的查看和分析，形成统一的不间断的数据服务，带来了更直观、更深入的数据体验方式。

6.6.6 双创大数据版图的意义

清数双创大数据版图通过展现资金、人才、园区等创业要素在行业和地区间的发展和流动情况，让观察者能深入剖析地区双创趋势，洞见技术和市场的发展浪潮；并特别提出清数双创指数来整体反映一个地区的双创活跃度，为政府和企业决策提供重要的参考依据。目

前清数中国创新创业大数据版图已经正式发布，来自四川、重庆、云南、江苏、湖南、西藏等地的政府领导、投资机构和创业者在参观后都表示出了强烈的兴趣，一致认为清数发布的双创大数据版图将成为双创形势的基本指南。清数中国创新创业大数据版图，可以为政府、企业、创业者提供全面的价值参考，帮助实现双创数据价值的转化。对于政府、企业可以促进决策优化，从而带来招商引资和产业发展的机会，对于创业者可以时刻关注创业动态，从版图中发现价值洼地和成长机遇。

第二部分　数据科学和数据工程

数据科学和数据工程共分为7章，主要内容有：数据科学概念、研究重要角色、生命周期管理、数据仓库、数据挖掘分析方法、知识发现及大数据处理平台，通过建立科学系统的数据分析方法论，指导数据工程实践；在数据工程方面，重点介绍医疗行业大数据、环保行业大数据、移动社交大数据、金融行业大数据和工业制造大数据等几个热点行业数据工程实践，每个行业又侧重大数据应用的不同角度，总体上全面解析大数据应用的多个方面。最后提出大数据工程保障体系建设，包括法律体系建设、标准体系建设、标准化大数据治理体系建设、技术和应用研究、创新平台建设等。该部分章节充分体现了理论性、科学性、创新性、实用性、经济性、社会性、标准性、保障性和完整性，形成数据科学和数据工程体系。



数据科学理论与工具

7.1 数据科学理论基础

知识经济(基于知识的资本)中知识的增长与知识的数字化基本上是同步的。在2012年年初达沃斯世界经济论坛上,一份题为《大数据,大影响》的报告宣称,数据已成为一种新的经济资产类别。那么一个很自然的推论是,数据的贡献就应该被合理地计量。然而目前传统的经济统计方法测量的对象主要是商品和服务,并不能很好地适应于数据。Mandel(2012)认为,在数据驱动经济的框架下,各种数字信息的生产、分配和使用是驱动经济增长的重要因素,而经济增长、消费、投资和贸易等宏观指标的测量低估了数据的贡献。已故图灵奖得主格雷(Jim Gray)在20世纪90年代中期曾指出,数据库技术的下一个"大数据"挑战将会来自科学领域而非商业领域,并且提出了科学研究的第四范式是数据密集型科学。在《大数据时代的历史机遇:产业变革与数据科学》(2013)一书中,鄂维南院士也提到:"大数据在科学领域的表现是数据科学的兴起,数据科学将成为科研体系中的重要组成部分,并逐渐达到与物理、化学、生命科学等自然科学分庭抗争的地位。"然而数据科学目前只是多个相关学科"拼接"起来的一个新兴学科,尚未形成完整的学科框架体系。

7.1.1 数据科学概念

大数据的热潮,催生了一门新的学科即数据科学。数据科学正处于发展初期,是一门不断发展的学科。数据科学的核心涉及用自动化的方法来分析海量数据,并从中提取知识。在几乎所有的知识发现领域,数据科学提供了一种强大的新方法来探索发现,它为拥有大量数据但不知怎样从数据中提取价值的公司提供了一种新的见解来源。伴随着这种自动化方法的发展,数据科学正在帮助创造新的科学分支并影响着社会科学和人文科学领域。数据科学融合了多门学科并且建立在这些学科的理论和技术之上,包括数学、概率模型、统计学、机器学习、数据仓库、可视化等。在实际应用中,数据科学包括数据的收集、清洗、分析、可视化以及数据应用整个迭代过程,最终帮助组织制定正确的发展决策。数据科学的从业者称为数据科学家。

数据科学目前还没有明确的基础理论,人们对数据科学的定义各不相同。许多学者立足各自的视角对数据科学的基础理论提出了不同的观点,例如,V. Dhar将数据科学定义为研究从数据中提取知识的一门学科。J. Leak认为数据科学其关键词是"科学"而不是"数据"。复旦大学数据科学研究中心的朱扬勇教授则认为数据科学是关于数据的科学或者研究数据的科学,定义为:研究探索Cyberspace中数据界奥秘的理论、方法和技术,研究的对

象是数据界中的数据。因此，数据科学要作为一门独立的学科存在，还需要更多的学术认同和大量长期的实践积累。

数据科学的广义定义为研究探索 Cyberspace 中数据界（datanature）奥秘的理论、方法和技术，研究的对象是数据界中的数据。数据科学的研究对象是 Cyberspace 的数据，是新的科学。数据科学主要有两个内涵：一个是研究数据本身，研究数据的各种类型、状态、属性及变化形式和变化规律；另一个是为自然科学和社会科学研究提供一种新的方法，称为科学研究的数据方法，其目的在于揭示自然界和人类行为现象和规律。狭义定义为数据科学是研究数据的科学。它利用统计学知识和计算机技术对专业领域的对象进行现实大数据分析与挖掘及其他方式的数据处理，以使组织获取更大的经济效益。

目前，学者们从不同角度对数据科学给出了一种定义。数据科学是一门将“现实世界”映射到“数据世界”之后，在“数据层次上”研究“现实世界”的问题，并根据“数据世界”的分析结果，对“现实世界”进行预测、洞见、解释或决策的新兴科学；是以“数据”尤其是“大数据”为研究对象，并以数据统计、机器学习、数据可视化等为理论基础，主要研究数据预处理、数据管理、数据计算等活动的交叉性学科；是以实现“从数据到信息”“从数据到知识”和（或）“从数据到智慧”的转化为主要研究目的，以“数据驱动”“数据业务化”“数据洞见”“数据产品研发”和（或）“数据生态系统建设”为主要研究任务的独立学科；是以“数据时代”尤其是“大数据时代”面临的新挑战、新机会、新思维和新方法为核心内容的，包括新的理论、方法、模型、技术、平台、工具、应用和最佳实践在内的一整套知识体系。

大数据（以半/非结构型数据为主）使基于关系型数据库的传统分析工具很难发挥作用，或者说传统的数据库和统计分析方法很难在可容忍的时间范围内完成存储、管理和分析等一系列数据处理过程，为了有效地处理这类数据，需要一种新的范式——数据科学。真正意义上的现代统计学是从处理小数据、不完美的实验等这类现实问题发展起来的，而数据科学是因为处理大数据这类现实问题而兴起的。因此数据科学的研究对象是大数据，而统计学以结构型数据为研究对象。退一步，单从数量级来讲，也已发生了质变。对于结构化的大规模数据，传统的方法只是理论上的（可行性）或不经济的（有效性），实践中还需要借助数据挖掘、机器学习、并行处理技术等现代计算技术才能实现。

7.1.2 数据科学预测预警分析

调查发现，如今有超过一半的企业领导认识到他们无法获取完成自己的工作所需的数据（Paul C. Zikopoulos，Chris Eaton，Dirk de Roos，Thomas Deutsch，George Lapis 所著 *Understanding Big Data*）。企业的数据资产以滚雪球似的速度增长，尽管无论从硬件设施还是软件技术，企业都有能力存储这些数据，但是从海量的、多样的和实时增长的数据资产中挖掘“金矿”，为企业提供精准的商业洞察，提升服务水平和提高商业价值，是企业所面临的挑战。

预测是在时间序列和周期运行基础上识别模式进而在相似情景下外推、应用模式的过程。MIT 研究的显示，人们 93%的行为可以预测。我知道我两个月后的周六上午 10:00，在点评“网络的效应”，或者“跨国分层网络模型如何建立”的研讨班主题的概率超过 90%。而人是一个很强大的模式识别机器，大数据集成 5V 数据，可以帮助寻找、分析和发现模式；这是 ICT 技术对于人脑的高级模仿，因为人脑就是一个典型的大数据处理装置，IBM

Watson 以其在“危险边缘”节目的出色表现，再次提供了人类与机器模型相互模仿、相互学习、相互协同的典型案例。发现模式是预测的基础，大数据加上人工智能算法可以强化“类人”的模式识别能力；二十年前，有了数据库和数据挖掘，人们对于啤酒与尿布的例子津津乐道，而如今有了大数据，或许它会帮助我们厘清“教授、啤酒炸鸡、韩剧、世界杯”之间的关系，可以洞察球迷的行为，以及辨别真球迷和假球迷、伪球迷的区别。

预测性分析涵盖了各种统计学技术，包括利用预测模型、机器学习、数据挖掘等技术来分析当前及历史数据，从而对未来或其他不确定的事件进行预测。在商业领域，预测模型从历史数据探索规律，以识别可能的风险和商机。预测性分析是数据挖掘技术的延伸技术，它主要用来对未来情况进行预测，以帮助决策层做出更加正确的决定。

(1) 流数据分析为业务提供实时决策响应。流数据是那些随着时间推移而无限增长的数据集合，流数据分析用于识别数据流中的复杂事件，并提供实时的分析报告及决策相应。例如，使用回归模型分析购买的交易数据，分析用户购买行为，从而判定欺诈性消费行为，规避用户损失。

(2) 非结构化数据的预测分析，探索新的业务价值。常见的非结构化数据包括 Web 日志、企业知识库、网络文字、图形、视频及声频等更加难以解析的数据源，非结构化数据分析用于从非结构化的数据源中通过语义分析和词法分析技术提取关键词，并采用聚类、关联或其他算法预测分析，探索新的业务价值。例如，舆情分析通过文本挖掘提取结构化客户舆情及“用户”，并通过算法分析，识别联系人、关注者及其关系，形成社交网络。

模式识别是指通过计算机用数学技术方法来研究模式的自动处理和判读。我们把环境与客体统称为“模式”。随着计算机技术的发展，人类有可能研究复杂的信息处理过程。信息处理过程的一个重要形式是生命体对环境及客体的识别。对人类来说，特别重要的是对光学信息(通过视觉器官来获得)和声学信息(通过听觉器官来获得)的识别。这是模式识别的两个重要方面。

(3) 结构化数据的深度挖掘，深入剖析业务价值。结构化数据是数据仓库或其他操作性数据库中的数据，这些数据用于传统 OLAP 和商业智能等，生产的报告用于说明发生了什么，以及了解过去和现状。但是随着业务需求的发展，我们更需要知道为什么发生，将来会发生什么，甚至最佳结果是什么，用于揭示隐藏的关联关系及趋势，这就需要我们对这些数据进行深度的挖掘，生成预测模型。例如，可以通过预测分析，提供精准营销，促进企业发展。

7.1.3　商业智能与数据科学

每当提及“数据科学”，人们总是会联想到另外一个含义似乎类似，却又无法清楚区分的名词——商业智能(Business Intelligence，BI)。在此有必要通过对比来区分这两个概念。

商业智能致力于使用一组统一的衡量标准来评估企业过去的绩效指标，并用于后续的业务规划。这包括建立关键绩效指标(Key Performance Indicator，KPI)，用于表示评估业务的最基本的衡量标准。测量尺度和关键绩效指标通常都是在联机分析处理模式(OLAP Schema)中定义，使得商业智能报表的内容能够基于已定义的衡量标准。

商业智能的典型技术和数据类型包括：

(1) 标准和满足特定需求的报表、信息面板、警报、查询及细节；

(2) 解构化数据、传统数据源、易操作的数据集。

数据科学可以简单地理解为预测分析和数据挖掘，是统计分析和机器学习技术的结合，用于获取数据中的推断和洞察力。相关方法包括回归分析、关联规则(比如市场购物篮分析)、优化技术和仿真(比如蒙特卡罗仿真用于构建场景结果)。

数据科学的典型技术和数据类型包括：

(1) 优化模型、预测模型、预报、统计分析；

(2) 结构化/非结构化数据、多种类型数据源、超大数据集。

商业智能和数据科学都是企业所需要的，用于应对不断出现的各种商业挑战。如图 7-1 所示，展示了商业智能和数据科学的不同定位和范畴。由图可以看出，商业智能更关注于过去的旧数据，其结果的商业价值相对较低；而数据科学更着眼于新数据和对未来的预测，其商业价值相对较高。但是我们也看到，这两个区域使用虚线分割。换言之，它们并不存在一个明确的划分，只是各有偏重而已。

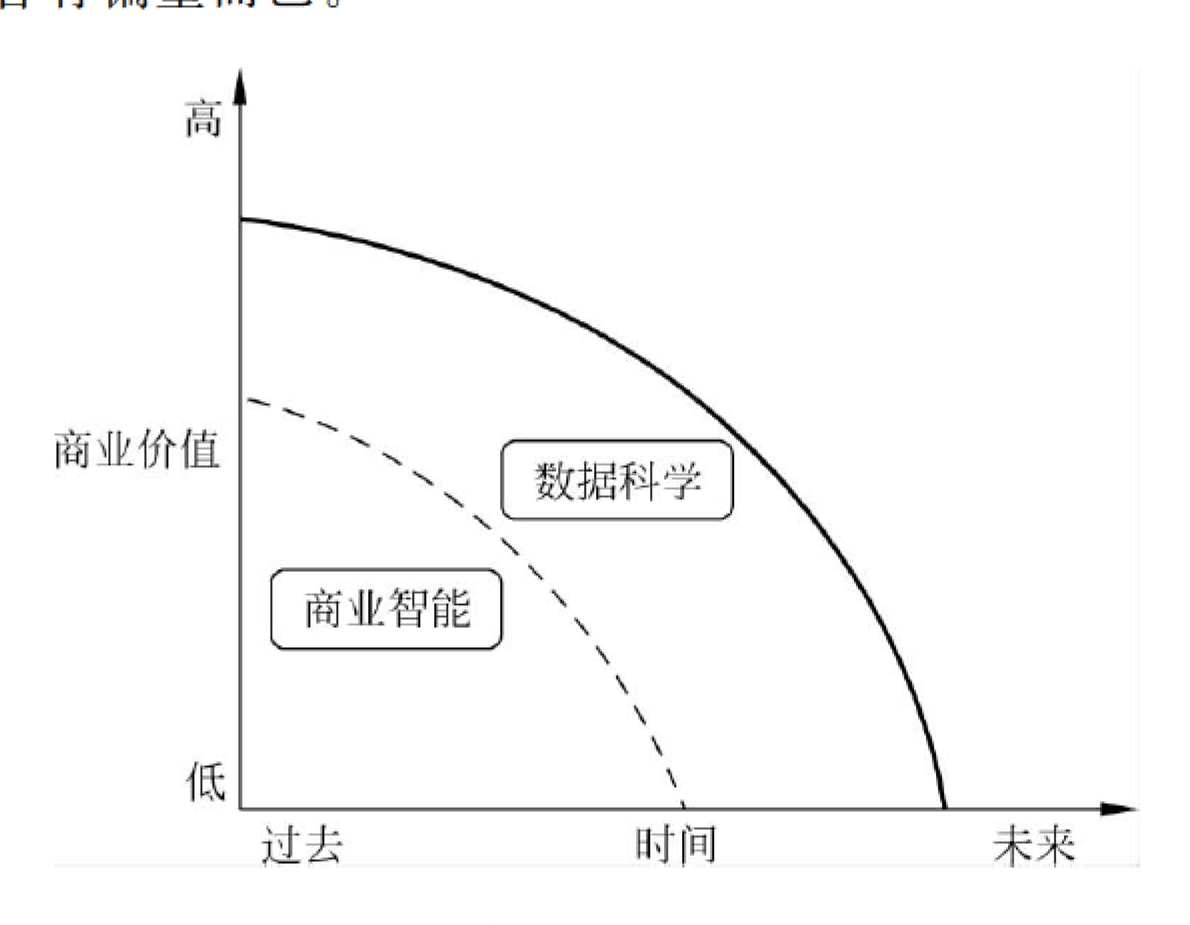


图 7-1　商业智能与数据科学

数据科学是大数据发展的理论支持，不仅要解决大数据的存储和管理，还要实现预测式分析。数据科学是统计学的论证，真正利用到统计学的力量。只有这样才能够从数据中获得经验和未来方向的指导。但是，数据科学并非简单的统计学，需要新的应用、新的平台和新的数据观，而不仅是现有的传统的基础架构与软件平台。

7.2　数据科学研究的重要角色

大数据的出现，催生了新的数据生态系统。为了提供有效的数据服务，它需要三种典型的角色。表 7-1 介绍了这三种角色，以及每种角色具有代表性的专业人员举例。

表 7-1　数据科学研究重要角色

角　色	描　述	专业人员举例
深度分析人才	通过定量学科(例如数学、统计学、机器学习)高等训练的人员；精通技术，具有非常强的分析技能和处理原始数据、非结构化数据的综合能力，熟悉大规模复杂分析技术	数据科学家、统计学家、经济学家、数学家

续表

角　　色	描　　述	专业人员举例
数据理解专业人员	具有统计学和/或机器学习基本知识的人员：知道如何定义使用先进分析方法可以解决的关键问题	金融分析师、市场研究分析师、生命科学家、运维经理、业务和职能经理
技术和数据的使用者	提供专业技术用于支持分析型项目的人员：技能包括计算机程序设计和数据库管理	计算机程序员、数据库管理员、计算机系统分析师

典型的分析型项目需要多种角色。但数据科学家是自身结合了多种以前被分离的技能，成为一个单一的角色。以前是不同的人用于一个项目的各个方面，比如，有的人去应对业务线上的终端用户，另外的具有技术和定量专长的人去解决分析问题。数据科学家就是这些方面的综合体，有助于提供连续性的分析过程。

7.2.1　数据科学家

数据科学家（Data Scientist）能够提供用于分析技术、数据建模的学科专业知识，针对给定的业务问题使用有效的分析技术，并确保达到整体分析目标。数据科学家是有着开阔视野的复合型人才，他们既有坚实的数据科学基础，如数学、统计学、计算机学等，又具备广泛的业务知识和经验。数据科学家通过精深的技术和专业知识在某些学科领域解决复杂的数据问题，从而制定出适合不同决策人员的大数据计划和策略。数据科学家负责为复杂的业务问题建模、发现业务洞察力并找到新的商业机遇。

1. 数据科学家需要具备的主要能力和行为特征

(1) 定量技能：比如数学或统计学技能。

(2) 技术才能：比如软件工程、机器学习和编程能力。

(3) 善于怀疑：对于数据科学家来说，能够采用批判的眼光来审视自己的工作，而不是采用片面的求同方式，这是很重要的。

(4) 好奇心和创造力：数据科学家必须对数据充满激情，并能够找到创新的方式来解决问题和描述信息。

(5) 善于沟通和合作：即使具有很强的定量和工程技能也是不够的。数据科学家必须能够采用清晰的方式表达出项目中的商业价值，并能与项目发起人（Sponsor）和项目干系人（Stakeholder）合作工作，从而让其在项目中产生共鸣。

2. 数据科学家的关键活动

(1) 将商业挑战构建成数据分析问题；

(2) 在大数据上设计、实现和部署统计模型和数据挖掘方法；

(3) 获取有利于引领可操作建议的洞察力。

寻找技能熟练的人才是与大数据分析相关的主要挑战之一。成功的大数据分析计划要求 IT 部门、业务用户等众多关键角色和数据科学家之间紧密协作，以选择和实施可以正确解决业务问题的分析。

7.2.2 数据科学与工程相关角色

通常在项目中会有各种角色和主要项目干系人。每个角色都在分析型项目中起到各自不同的作用。如图 7-2 所示列出了 7 个角色，实际中可根据项目的工作范围、组织结构和参与者的技能要求，选择适当的项目参与角色人选。

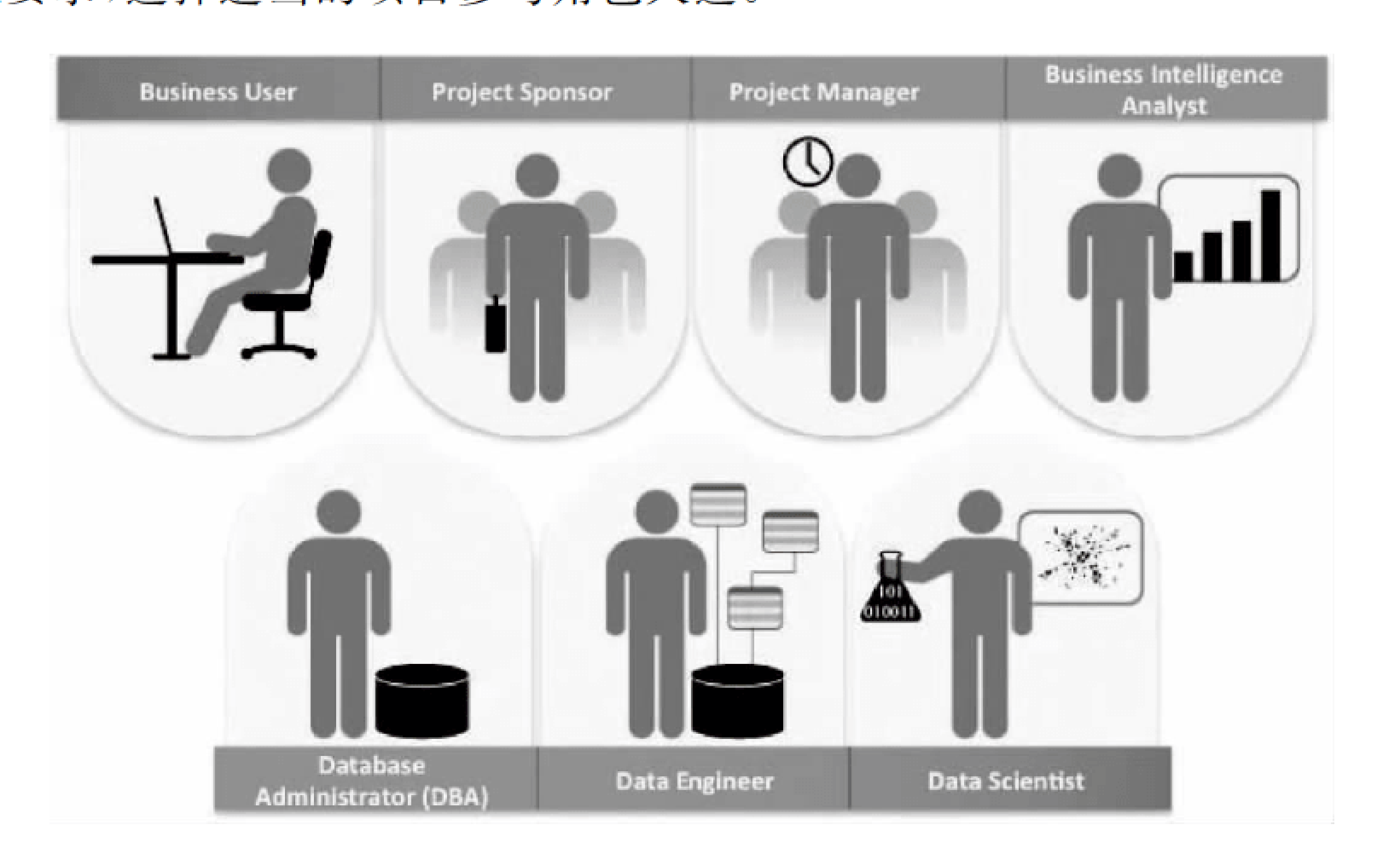


图 7-2 数据科学与工程中的关键角色

(1) 业务用户(Business User)：那些从最终结果中受益的人。可以充当项目团队的顾问，并提供建议，例如，如何评估最终结果的价值和如何实施它们。

(2) 项目发起人(Project Sponsor)：负责项目的启动。提供项目的推动力和核心业务问题。通常提供资金，并判定来自工作团队的最终输出的价值的程度。

(3) 项目经理(Project Manager)：确保以期待的质量按时达到关键里程碑和目标。

(4) 商业智能分析师(Business Intelligence Analyst)：提供业务领域专业知识，对数据、关键绩效指标、关键衡量标准和报表视角的商业智能有着深入的理解。

(5) 数据工程师(Data Engineer)：利用深入的技术技能，协助优化 SQL 查询，用于数据管理、提取和支持分析沙盒摄取数据。

(6) 数据库管理员(Data Base Administrator，DBA)：提供和配置数据库环境，用以支持工作团队的分析需求。

(7) 数据科学家(Data Scientist)：提供用于分析技术、数据建模的学科专业知识，针对给定的业务问题使用有效的分析技术，并确保达到整体分析目标。

数据科学与工程是一个新兴领域，同时数据科学家是拥有特殊技能的全新专业人员。数据科学家负责为复杂的业务问题建模、发现业务洞察力并找到新的商业机遇。

7.3 大数据生命周期管理方法论

首先，需要了解一下使用这样一个数据分析生命周期模型的价值何在。很多问题看上去相当复杂难解，但是一个定义良好的流程能够帮助数据科学家将复杂的问题分解成更容易处理的过程。使用一个好的流程进行分析是极其重要的，因为它既有助于实现全面可重

复实施的分析方法，又可以让数据科学家把必要的精力尽早地放到那些可以掌握问题重点的步骤中。人们经常不愿意花太多时间去做大量的计划、调研或者问题解决等工作，而是急于开始收集和分析数据。这样做很可能出现的结果是：项目成员在中途发现正在尝试解决的问题和项目发起人的目的截然不同或者与之前的沟通结果不一样。创建并文档化一个流程将有助于展示项目的分析结果的严谨性。当我们谈及发现的结果时，这将为项目提供额外的可信度。这个流程还使我们能够去教别人如何使用这些方法和分析，以使得它是可以在下个季度、下一年或者被新的员工重复使用的。

虽然一个定义良好的流程有助于指导我们完成任何一个分析项目，然而需要特别声明的是：本节将着重介绍的数据分析生命周期模型更适合数据科学项目。与着眼于获取关键绩效指标或者实现信息系统其他功能项目相比，数据科学项目还是会有些相似的步骤。

7.3.1 数据分析模型概述

数据分析生命周期模型描述了一种针对端到端分析过程(从商业理解到项目完成)的最佳实践方法。并且，其中一些用于改进模型的步骤来源于数据分析和决策科学范畴中已有的方法。这些已有的方法提供了流程中的一部分或者使用不同术语的类似概念。一些参考过的流程如下。

科学方法(Scientific Method)，虽然已经有上百年历史了，但是依然提供了严谨的框架，用于思考问题并将其解构为多个主要部分。

跨行业数据挖掘标准流程(Cross Industry Standard Process for Data Mining，CRISP-DM)提供了一些有用的考虑分析型问题的方法。

Tom Davenport 在他的《工作中的分析》一书中提出的 DELTA 框架。

Doug Hubbard 的应用信息经济学(Applied Information Economics，AIE)方法。

MAD(Magnetic，Agile，Deep)涉及数据分析生命周期模型里阶段 4 至阶段 5 中关注的建立模型和模型评估。

目前，已经有很多成熟的方法模型，为数据挖掘实际应用提供了指导模型，其中，CRISP-DM(Cross-industry Standard Process for Data Mining，跨行业数据挖掘标准流程)，为 20 世纪 90 年代由全球领先的数据挖掘专家 SPSS 联合 NCR 公司、戴姆勒-克莱斯勒以及 OHRA 共同推出的全球首个数据挖掘行业方法论，并成立了 CRISP-DM 专家组(Special Interest Group，SIG)。SIG 拥有来自世界各地的两百多名成员，并获得了来自广泛领域内对数据挖掘感兴趣的从业者的帮助，包括数据集的提供者和管理顾问。从技术原理来讲，CRISP-DM 还不是一个成熟的理论，还没有形成一个实践性的、成功的、被广泛采纳的标准。CRISP-DM 的专家组 SIG 一直在为建立 CRISP-DM 方法学而努力，不断地在数据挖掘项目的实践当中积累经验，他们想要建立一个跨行业的公开的数据挖掘标准，并不断发展 CRISP-DM。

CRISP-DM 方法论的推出确实是及时的、有价值的，其设计背后有着广泛的经验支持，以保证该模型能够适应任何数据挖掘应用，包括欺诈发现、信用风险评估、税户保持、流失分析和税户赢回。数据分析生命周期由 6 个阶段组成。如图 7-3 所示，展示了这一数据挖掘过程的各个阶段，这些阶段之间的顺序并不固定，在不同阶段之间来回反复往往是非常有必要的。究竟下一步要执行哪个阶段或者哪一个特定的任务，都取决于每一个阶段的结果。

图中的箭头表明了阶段之间最重要和最频繁的依赖关系。图中最外层的这个循环表明了数据挖掘本身的循环性质。经过一个具体的数据挖掘项目得到了某项解决措施或方法并加以展开,并不代表数据挖掘本身已经结束。从这一数据挖掘过程以及解决措施展开的过程中所吸取的经验、教训,又引发了新的,通常是更加焦点的商业问题。接下来的数据挖掘过程将会从过去的项目经验中获利。

CRISP-DM 模型定义了 6 个过程,分别是:商业理解、数据理解、数据准备、建立模型、模型评估、结果发布。

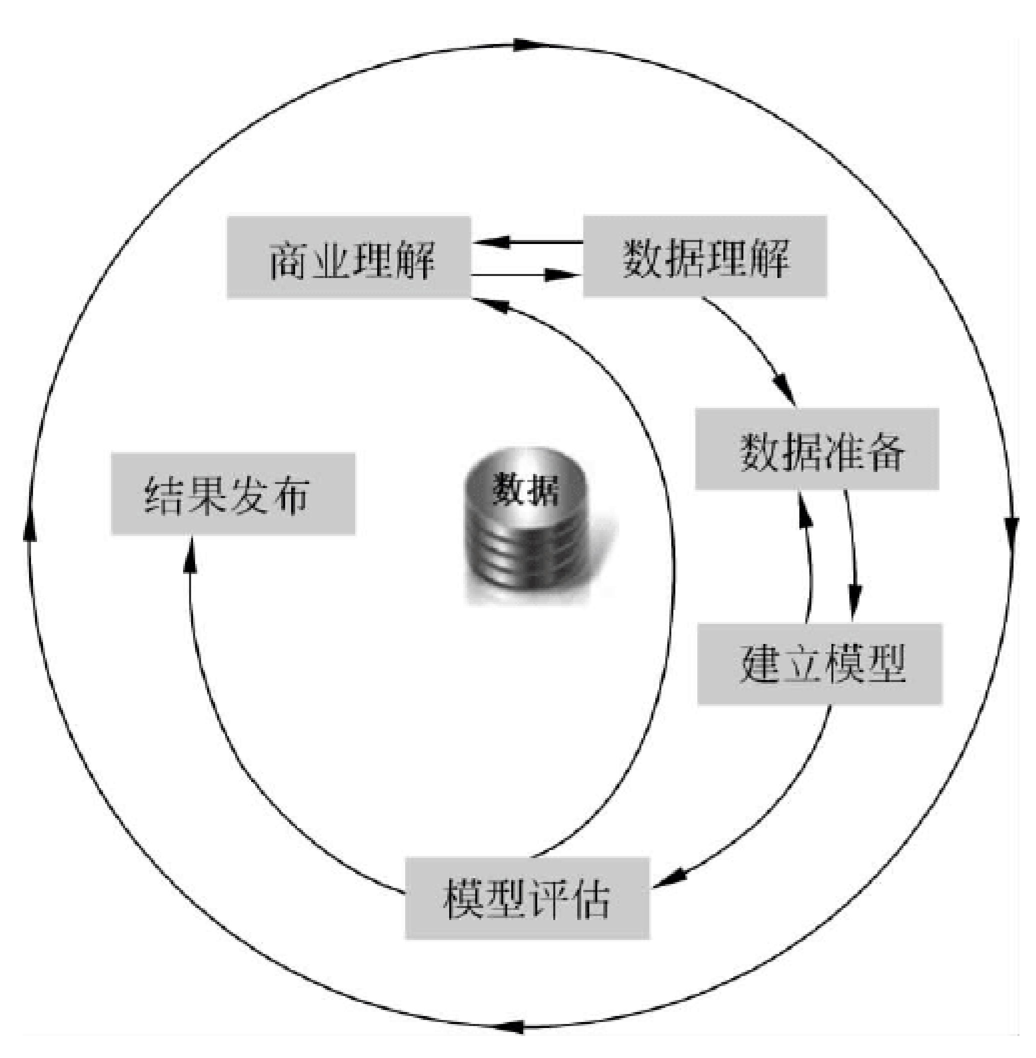


图 7-3 大数据分析生命周期模型

下面主要介绍常用的数据挖掘的生命周期模型 CRISP-DM,主要包括 6 个过程(www.crisp-dm.org)。

1. 商业理解

商业理解阶段的主要任务是理解项目的背景,即商业愿景和商业目标,把要实现的商业目标转化为相对应的数据挖掘的问题,并制定完成目标的工作计划。从业务的角度上,了解项目的要求和最终的目的是什么,并将这些目的与数据挖掘的定义和结果结合起来。

2. 数据理解

数据理解阶段是着手对源数据的收集,鉴别数据的质量问题,从数据中发现隐藏的信息或探测臆想的数据子集。理解数据,包括数据规模、数据列的特性等,并对可用的数据进行评估。

3. 数据准备

数据准备阶段是对源数据进行采集、清洗、转换,以满足数据建模要求。在源数据的基础上运用建模工具建立最终的数据集。数据准备可能重复多次,其主要任务是使用建模工具来传输和清洗数据,包括表、记录和属性等。

数据预处理阶段覆盖了从最初的源数据构造最终数据集合的所有活动,最终的数据集将会作为数据挖掘模型的输入或者样本数据。数据预处理阶段的某个任务很可能要执行多次,并且这些任务的执行顺序并不是固定的。数据预处理包括表、记录、属性的选择,以及为了根据数据挖掘算法的特点和要求对数据进行的清洗、转换和整理。对于数据挖掘来说,数据质量对挖掘效果的影响非常大,甚至可能导致错误的预测结果。因此,在任何时候都不要

忽视数据的质量,一般的数据挖掘过程中大约有一半的时间用于数据预处理。

这一阶段包括的功能节点有选择属性集、异常数据处理、缺失值处理、数据标准化、数据类型转化、度量尺寸设置、增加新列、排序、加权处理、计数、分类汇总、数据分组、抽样、选择数据集等。这些功能节点基本上可以满足用户对数据进行预处理的需要。经过数据预处理,用户可以选择与数据挖掘目标相关的、高质量的、适当容量的数据,包括属性(列)和记录(行)的选择。在这个阶段,用户使用数据理解阶段得到的数据质量问题解决方案对数据进行处理。

4. 建立模型

建立模型阶段,多种建模技术被选择和应用,它们的参数被校对到最理想的数值。一些技术解决同样的数据挖掘问题,一些技术需要特定的数据格式,因此建立模型阶段有时也需要重新进行数据准备。通过数据挖掘算法建立挖掘模型。通常,通过设置参数运行模型,再对这些参数进行微调或回到数据准备阶段以便执行所选模型所需要的操作,建模时通常会执行多次迭代,才会达到最终效果。

5. 模型评估

模型评估阶段,将从数据分析的观点建立一个或一些高质量的模型。在配置这些模型前,最重要的就是对已经建立的模型进行彻底的评估,并回顾建造模型的每一个步骤,确定商业目标被完全地达到。关键目标是确定一些重要的商业问题是否被充分地考虑,最终决定数据挖掘结果的使用目的是否达到。对建立的模型进行评估,根据在商业理解中定义的挖掘目标进行评定,以确保满足业务需求。

6. 结果发布

结果发布阶段,根据用户需要可能只是简单地创建一个报表,也可能是实现一个重复的、复杂的数据挖掘过程。在大多数的情况下,模型应该由用户,而不是数据分析师来配置。然而即使分析师不配置模型,对他来说重要的是让用户预先理解所要执行的配置动作,目的是让用户使用创建的模型。数据挖掘的结果和过程发布成可读文本形式,并通过数据挖掘结果进行改善。

7.3.2　数据分析模型流程框架

数据分析流程框架也就是数据分析生命周期模型(Data Analytics Lifecycle),如图 7-4 所示。该框架描述了数据分析生命周期模型各个阶段的流程执行和工作任务。

7.3.3　数据分析模型创新案例

以下结合具体的案例——全球创新网络分析(Global Innovative Network and Analytics, GINA)详细分析数据分析模型。该项目致力于分析企业内全球范围的创新活动数据,帮助数据科学家理解这些创新活动的深层含义,从而促进全球范围的创新活动更有借鉴意义。GINA 正是遵循数据模型分析框架展开的,接下来详细介绍 GINA 在各个分析阶段所进行的活动。

全球创新网络和分析(GINA)团队是一组位于世界各地高级技术专家的卓越中心(COE)。这个团队的章程是让员工跨越全球 COEs 来推动创新、研究和大学合作。2012 年,一

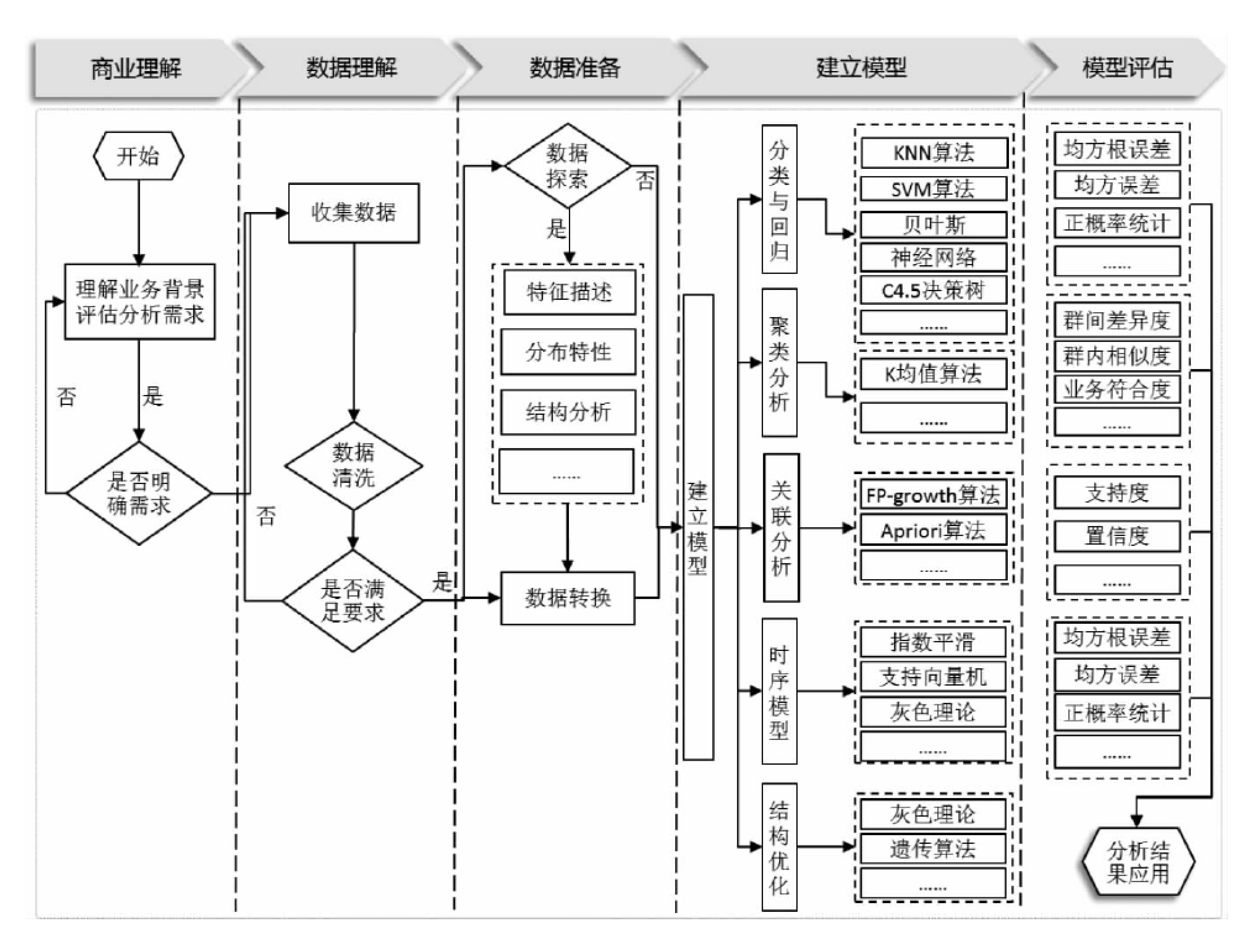


图 7-4　数据分析模型框架

个新聘的首席信息官想改进这些活动，并提供一种跟踪和分析相关信息的机制。此外，这个团队希望创建更强大的机制来捕获其非正式对话的结果与学术界或其他组织中的其他思想领袖，从中进行数据挖掘以获得知识发现。

1. 商业理解

在 GINA 项目的开始阶段，团队要识别数据源。虽然 GINA 团队成员对工程技术方面较熟悉，但还是有一些数据和想法需要探索，还缺乏一个正式的团队来执行这些分析。在咨询了包括巴布森学院(Babson College)的知名分析专家 Tom Davenport、麻省理工学院集体智慧专家兼协同创新网络(Collaborative Innovation Networks，CoIN)创始人 Peter Gloor 等专家后，团队决定在全球范围内寻找志愿者来推广工作。

GINA 团队认为将组织全球团队成员来分享想法，并能够实现知识共享。它计划创建一个数据存储库包含结构化和非结构化数据，主要实现以下三个目标。

(1) 存储正式和非正式数据。

(2) 跟踪全球技术专家的研究。

(3) 挖掘数据的模式和洞察，以改善团队的运营和战略。

GINA 案例研究提供了一个团队如何应用数据分析生命周期进行分析的示例。创新通常是一个难以衡量的概念，这个团队想通过使用先进的分析方法来确定公司内部优秀的创新者。

这里是一个工作团队的各种角色的分工情况。

(1) 业务用户，项目赞助商，项目经理：CTO 办公室副主席。

（2）商业智能分析师：IT代表。

（3）数据工程师和数据库管理员（DBA）：IT代表。

（4）数据科学家：杰出工程师，也开发了GINA中显示的社会图表案例分析。

GINA团队开发的10个主要假设如下：

假设1：在不同地理区域的创新活动可以映射到企业的战略方向。

假设2：当全球知识转移作为构想交付过程的一部分时，交付想法所需的时间长度减少。

假设3：参与全球知识转移的创新者比不参与全球知识转移的创新者更快地提供想法。

假设4：可以分析和评估创新者提交的意见是否有可能获得资金。

假设5：可以跨地理区域测量和比较特定主题的知识发现和增长。

假设6：知识转移活动可以识别不同地区的研究特定超长者。

假设7：战略性的企业主题可以映射到地理区域。

假设8：频繁的知识扩展和转移事件减少了从想法生成公司资产所需的时间。

假设9：使用谱系图可以揭示知识扩展和转移产生公司资产的情况。

假设10：新兴研究主题可以分类并映射到特定的创意者、创新者、超长者和资产。

GINA（IHs）可以分为两类：

（1）对当前发生的事情进行描述性分析，以进一步创造、协作和资产生成。

（2）预测分析，为执行管理层建议未来投资的地方。

2. 数据理解

项目发起的方法是利用社交媒体和博客加速全球创新和研究数据的收集，并激励全球各地的“志愿者”数据科学家团队。首先要组建一个项目团队，成员要满足有能力且有充分时间去处理那些复杂问题。数据科学家往往对数据充满激情，项目发起人能够利用这些热情的人才，以创造性的方式完成具有挑战性的工作。

该项目的数据分为以下两大类。

第一类是由发起公司的内部创新竞赛提出的五年的想法，被称为创新路线图（以前称为创新展示）。创新路线图是一个正式的、有机的创新过程，来自全球的员工提交的想法，然后审查和判断。选择最好的想法进一步孵化。因此，数据是结构化数据的混合，例如想法计数、提交日期；发明人姓名和非结构化内容，例如想法本身的文本描述。

第二类数据包括来自世界各地的创新和研究活动的时间和注释。这也表示结构化和非结构化数据的混合。结构化数据包括诸如日期、名称和地理位置的属性。非结构化文档包含表示公司内知识增长和转移的丰富数据的“who，what，when，where”信息。这种类型的信息通常存储在各个不同研究团队中几乎没有可见性的业务孤岛中。

3. 数据准备

团队与其IT部门合作，建立一个新的分析沙盒来存储和实验数据。在数据探索过程中，数据科学家和数据工程师开始注意到某些数据需要调节和正常化。此外，团队意识到几个失踪数据集对于测试一些分析假设至关重要。

当团队探索数据时，他们很快就意识到，如果数据的质量不够好或者没有足够的高质量数据，就无法执行生命周期过程中的后续步骤。因此，确定项目需要什么级别的数据质量和

清洁度非常重要。在GINA案例中,团队发现许多研究者和大学人员的名字被拼错,或者在数据存储中的首尾有空格。这些看似数据中的小问题都必须在本阶段解决,以便在随后阶段更好地分析和聚合数据。

为了达到模型的输入数据要求,需要对数据进行转换,包括生成衍生变量、一致化、标准化等。

1）数据清洗

现实世界的数据一般是不完整的、有噪声的和不一致的。数据清理例程试图填充缺失的值,光滑噪声并识别离群点,纠正数据中的不一致。

(1）缺失值处理。①忽略元组：当缺少类标号时通常这样做。除非元组有多个属性缺少值,否则该方法不是很有效。②人工填写缺失值：一般情况下,该方法很费时。③使用一个全局常量填充缺失值：将缺失值用同一个常数(如Unknown或−∞)替换。如果缺失值都用Unknown替换,则挖掘程序可能误认为它们形成了一个有趣的概念,因为它们都具有相同的值"Unknown"。因此此方法虽然简单但不可靠。④使用属性的均值填充缺失值。例如,假定顾客的平均收入为56 000美元,则使用该值替换income中的缺失值。⑤使用与给定元组属同一类的所有样本的属性均值,例如,将顾客按credit_risk分类,则用具有相同信用度给定元组的顾客的平均收入替换income中的缺失值。⑥使用最可能的值填充缺失值。可以用回归、使用贝叶斯形式化的基于推理的工具或决策树归纳确定。例如,利用数据集中其他顾客的属性,可以构造一棵决策树来预测income的缺失值。

(2）噪声数据处理。噪声是被测量的变量的随机误差或方差。给定一个数值属性(如price),怎样才能光滑数据,去掉噪声？下面介绍数据光滑技术。①分箱：分箱方法通过考察数据的"近邻"来光滑有序数据的值。有序值分布到一些桶或箱中。由于分箱方法考察近邻的值,因此是对数据进行局部光滑。例如,frequencly排序后数据(频次)：4,8,15,21,21,24,25,28,34划分为(等频)箱——箱1：4,8,15；箱2：21,21,24；箱3：25,28,34。用箱均值光滑：箱1：9,9,9；箱2：22,22,22；箱3：29,29,29。用箱边界光滑：箱1：4,4,15；箱2：21,21,24；箱3：25,25,34。②回归：可以用一个函数(如回归函数)拟合数据来光滑数据。③聚类：可以通过聚类检测离群点,将类似的值组织成群或簇。直观地,落在簇集合之外的值视为离群点。

(3）数据不一致的处理。作为一位数据分析人员,应当警惕编码使用的不一致问题和数据表示的不一致问题(如日期"2004/12/25"和"25/12/2004")。字段过载是另一种错误源,通常是由如下原因导致：开发者将新属性的定义挤压到已经定义的属性的未使用(位)部分(例如,使用一个属性未使用的位,该属性取值已经使用了32位中的31位)。

2）数据集成

数据分析任务多半涉及数据集成。数据集成是指将多个数据源中的数据合并并存放到一个一致的数据存储(如数据仓库)中。这些数据源可能包括多个数据库、数据立方体或一般文件。在数据集成时,有许多问题需要考虑。模式集成和对象匹配可能需要技巧。来自多个信息源的现实世界的等价实体如何才能匹配？这涉及实体识别问题。例如,数据分析者或计算机如何才能确信一个数据库中的innovator_id和另一个数据库中的inno_number指的是相同的属性？每个属性的元数据包括名字、含义、数据类型和属性的允许取值范围,以及处理空白、零或null值的空值规则。这样的元数据可以用来帮助避免模式集成的错

误。元数据还可以用来帮助变换数据(例如,pay type 的数据编码在一个数据库中可以是"H"和"S",而在另一个数据库中是 1 和 2)。因此,这一步也与前面介绍的数据清理有关。另外,冗余也是一个重要问题。一个属性可能是冗余的,如果它能由另一个或另一组属性导出。属性或维命名的不一致也可能导致结果数据集中的冗余。有些冗余可以被相关分析检测到。注意,相关并不意味因果关系。也就是说,如果 A 和 B 是相关的,这并不意味着 A 导致 B 或 B 导致 A。例如,在分析创新想法统计数据库时,可能发现一个地区的大学数与研究机构数是相关的,但这并不意味一个导致另一个。实际上,二者必然地关联到第三个属性:创新想法数。对于分类(离散)数据,两个属性 A 和 B 之间的相关联系可以通过卡方检验发现。除了检测属性间的冗余外,还应当在元组级检测重复。去规范化表的使用是数据冗余的另一个来源。数据集成的第三个重要问题是数据值冲突的检测与处理。例如,对于现实世界的同一实体,来自不同数据源的属性值可能不同。这可能是因为表示方法、比例或编码不同。例如,重量属性可能在一个系统中以公制单位存放,而在另一个系统中以英制单位存放。对于创新者,不同国家的收入可能涉及不同货币,而且可能涉及不同的福利和税。

3) 数据变换

数据变换是指将数据转换或统一成适合于挖掘的形式。

(1) 数据泛化:使用概念分层,用高层概念替换低层或"原始"数据。例如,分类的属性,如街道,可以泛化为较高层的概念,如城市或国家。类似地,数值属性如年龄,可以映射到较高层概念,如青年、中年和老年。

(2) 规范化:将属性数据按比例缩放,使之落入一个小的特定区间。大致可分为三种:最小最大规范化、z-score 规范化和按小数定标规范化。

(3) 属性构造:可以构造新的属性并添加到属性集中,以帮助挖掘过程。例如,可能希望根据属性 height 和 width 添加属性 area。通过属性构造可以发现关于数据属性间联系的丢失信息,这对知识发现是有用的。

4) 数据归约

(1) 数据立方体聚集:聚集操作用于数据立方体结构中的数据。

(2) 属性子集选择:通过删除不相关或冗余的属性(或维)减小数据集。属性子集选择的目标是找出最小属性集,使得数据类的概率分布尽可能地接近使用所有属性得到的原分布。对于属性子集选择,一般使用压缩搜索空间的启发式算法。通常,这些方法是贪心算法,在搜索属性空间时,总是做看上去当时最佳的选择。策略是做局部最优选择,期望由此导致全局最优解。在实践中,这种贪心算法是有效的,并可以逼近最优解。①逐步向前选择:该过程由空属性集作为归约集开始,确定原属性集中最好的属性,并将它添加到归约集中。在其后的每一次迭代步,将剩下的原属性集中最好的属性添加到该集合中。②逐步向后删除:该过程由整个属性集开始。在每一步,删除尚在属性集中最差的属性。③向前选择和向后删除的结合。④决策树归纳:决策树算法,如 ID3、C4.5 和 CART 最初是用于分类的。决策树归纳构造一个类似于流程图的结构,其中每个内部(非树叶)节点表示一个属性的测试,每个分支对应于测试的一个输出;每个外部(树叶)节点表示一个类预测。在每个节点,算法选择最好的属性,将数据划分成类。

(3) 维度归约:使用编码机制减小数据集的规模,例如,小波变换和主成分分析。

(4) 数值归约:用替代的、较小的数据表示替换或估计数据,如参数模型(只需要存放

模型参数，不是实际数据）或非参数方法，如聚类、抽样和使用直方图。

(5) 离散化和概念分层产生：属性的原始数据值用区间值或较高层的概念替换。数据离散化是一种数据归约形式，对于概念分层的自动产生是有用的。离散化和概念分层产生是数据挖掘强有力的工具，允许挖掘多个抽象层的数据。很重要的是，用于数据归约的计算时间不应当超过或“抵消”对归约数据挖掘节省的时间。

4. 建立模型

在 GINA 项目中，对于大多数数据集，使用社交网络分析技术似乎是可行的。在其他情况下，很难提出适当的由于缺乏数据，测试假设的方法。在一个案例（IH9）中，小组决定启动纵向研究开始跟踪关于人们发展新知识分子的数据点属性。这个数据收集将使团队能够测试未来的以下两个想法。

- IH8：频繁的知识扩展和传输事件减少了所需的时间从想法生成公司资产。
- IH9：沿袭地图可以揭示知识扩展和转移没有（或没有）结果在公司资产。

对于提出的纵向研究，团队需要为研究确立目标标准。具体来说，它需要确定一个成功的想法的终极目标已经穿越整个旅程。

与研究范围相关的参数包括以下考虑：

(1) 确定实现此目标的正确里程碑。

(2) 跟踪人们如何将想法从每个里程碑向目标移动。

(3) 使用几种不同的方法（取决于如何收集和组合数据）来比较时间和结果。这些可以是简单的 t 检验或可能涉及不同类型的分类算法。

在模型构建阶段，数据科学家们会综合考虑业务需求精度、数据情况、花费成本等因素，选择最合适的模型。通过知识转移活动的分析能识别出不同地区内特定研究方面的边界跨越者。下面将详细介绍边界跨越者识别的模型构建，构建过程中首先分析项目的目的，通过多个模型的事前假设和事后运行，然后通过后续的模型评估，进行优化、调整，以求建成最合适的模型。

(1) 寻找边界跨越者。在数据准备阶段，GINA 项目组将使用社交网络的数据来证明这一假设。图 7-5 就是数据科学家 John Cardente 生成的社交网络图，这是个有名的“爱尔兰蝴蝶”案例。这个模型的输入条件是来自世界各地的员工在企业想法展示竞赛提交的所有想法，每个圆圈代表了一个提交过想法的员工，圆圈之间的连线代表了两个员工参与一起提交了某个想法。圆圈的大小取决于该员工参与提交想法的数量，浅色的圆圈表示该员工提交的某个想法入围了当年竞赛的决赛阶段。

图 7-5 中有 5 个虚线圈出的部分要特别关注。GINA 项目组研究了这些虚线椭圆中的圆圈（代表想法提交者和参与者），发现这些椭圆的其中一个里，所有的人都是爱尔兰人。继续跟踪他们其中的几个人，发现他们在一起提出想法，源于他们都参与了企业在爱尔兰开展的有针对性的培训。

数据科学家 John Cardente 通过模型验证，并将企业想法展示竞赛活动中的数据以数值的形式通过可视化显示出来。为了验证企业想法展示竞赛活动数据中的知识转移活动能否识别出假说中提到的边界跨越者，John Cardente 做了进一步的可视化，如图 7-6 所示。

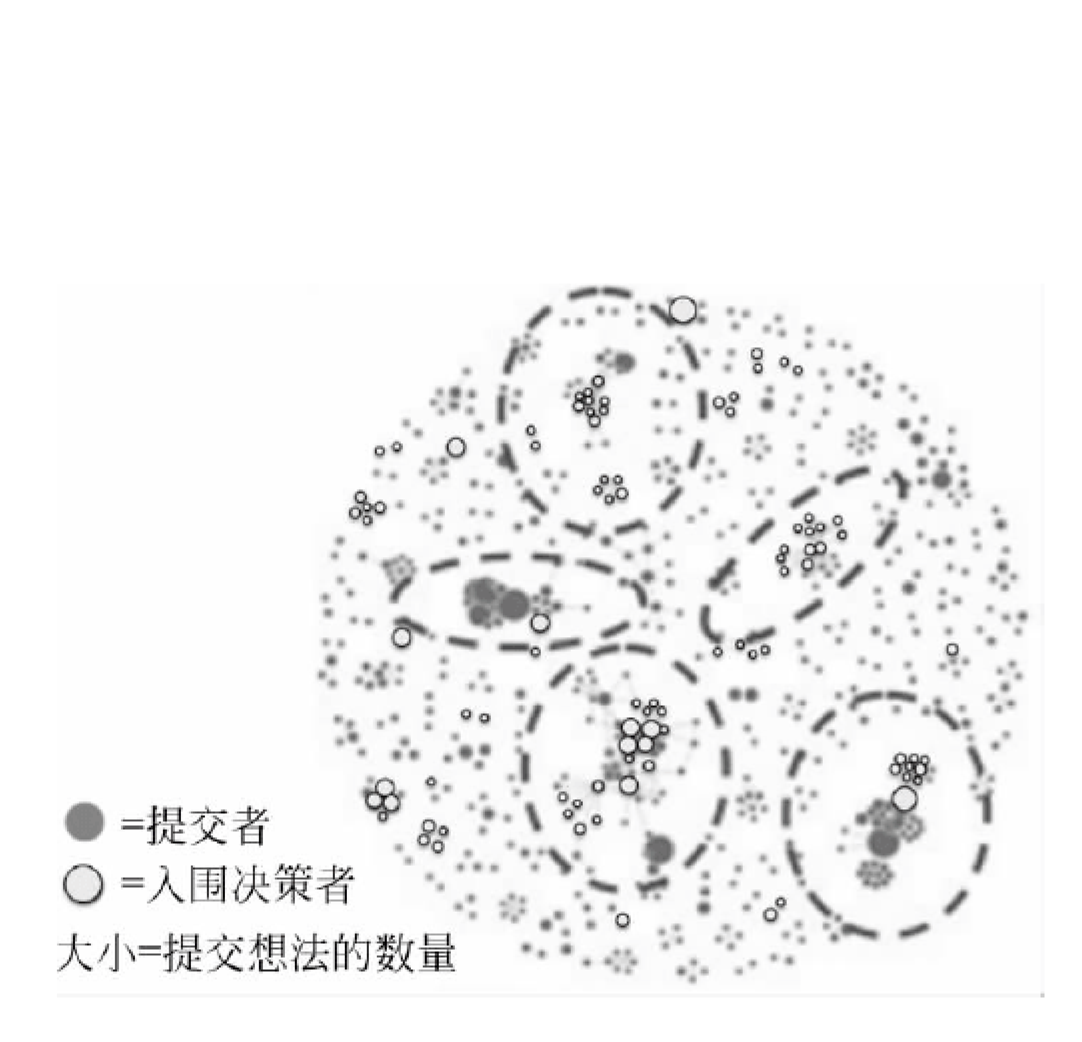


图 7-5 CINA 创新想法展示竞赛的社交网络图

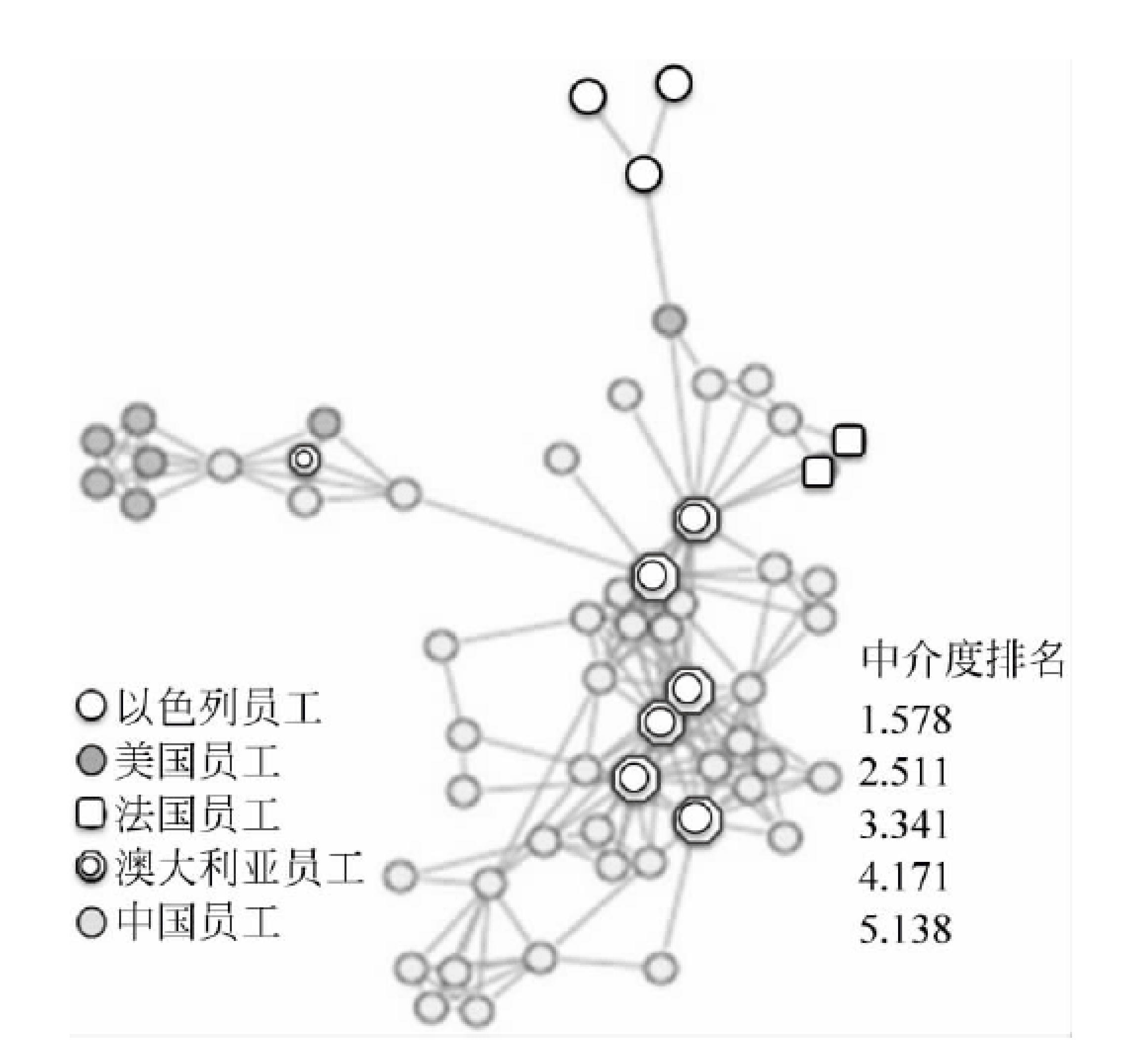


图 7-6 GINA 想法展示竞赛社交网络放大图

图 7-6 上面的白色圆形代表以色列员工，深色圆点代表美国员工。两个四边形圆点代表法国的员工。孤独的在左边的同心圆点代表来自澳大利亚的员工。浅色圆点是在整个实验中的最大集群之一，代表了中国的员工。

GINA 将具有白色圆圈的大圆点称为“枢纽”。这些人具有大量连线和很高的中介度(Betweenness)。在社会网络分析指标中，中介度用于衡量节点对于整个图连通性的重要度。在图 7-6 中，如果一些人中介度高，则说明这些人具有更大的影响力。图示中显示了中国的 5 大“中介度”的排名，五分之二中国员工的中介度得分远远高于其他地区。

(2) 验证创新卓越者。在图 7-6 中，有一名员工的中介度得分为 578，远远高于图中其他员工的中介度值，那么高中介度的他是不是就是创新卓越者呢？通过模型验证，输入他的名字查询结果如下。

① 2011 年，他参加了在希腊举行的 SIGMOD 会议。

② 他访问了企业在法国的业务部门，与一些员工见了面。

③ 他在一次自备餐会上讲述了他对 SIGMOD 会议自己的看法，这个会议的参与者包括：三名俄罗斯员工，一名埃及员工，一名爱尔兰员工，一名印度员工，三名美国员工和一名以色列员工。

④ 2012 年，他参加了在加州举行的 SDM 会议。

⑤ 在这次旅程中他拜访了 Greenplum 和 VMware 的创新者。

⑥ 在同一次旅程中，他参加了 CTO 评议会，将自己和其他两个研究员介绍给了数十个企业内的其他研究员。

这个结果跟我们假设的部分内容是基本吻合的，通过数据和分析模型 GINA 确实识别出了边界跨越者。

5. 模型评估

在第 6 阶段，GINA 团队发现了几种方法来剔除分析结果并确定最有影响力的结果和相关发现。这个项目被认为能成功识别边界跨越者。因此，数据科学家开展了纵向研究，开

始收集更长时间的数据并跟踪创新结果，不断完善数据和修改模型，以期达到研究目标。GINA 项目促进知识共享相关的创新和研究人员跨越公司内部和外部的多个领域，并且与大学建立关系联合研究关于数据科学和大数据。此外，该项目以有限的预算，利用高水平、杰出的工程师和数据科学家自愿完成。

该项目的一个重要发现是，在爱尔兰创新者密度不成比例。该项目在每年的创新大赛中，创新成果将为公司带来很大价值。当看 2011 年的数据时，15% 的获奖者来自爱尔兰。经过进一步研究，证明了爱尔兰的 COE 通过外部顾问接受了创新方面的重点培训。爱尔兰 COE 提出了比过去更多更好的创新想法。这在过去用传统的方法简直是不可能做到的。应用社交网络分析使得 GINA 团队能够在网络中找到一大批杰出的人才。这些发现是通过内部共享的演讲和会议，并通过社交媒体和博客等来推广的。

6. 结果发布

GINA 团队经过分析完整数据分析生命周期模型不难发现：在分析沙盒中对创新活动的各种会议纪要、记录和演讲文稿进行各种分析流程之后，我们可以对企业创新文化的深刻理解。项目进入第 7 阶段，GINA 将会把实验模型进行正式发布，先进行小规模生产系统部署试运行，同时熟悉和检验生产环境各方面的性能参数，以便在完全部署前做好各项准备。

项目的主要发现如下。

(1) GINA 未来需要更多数据，这就需要一个营销推广计划以说服员工提交在全球范围内他们进行的创新和研究活动。

(2) 这些数据是敏感的，团队需要考虑与数据相关的安全性和隐私性，例如，哪些人可以运行模型以及哪些人可以查看结果。

(3) 除了运行模型之外，还需要提供数据搜索功能。

(4) 部署后需要一种机制来不断评估模型。评估是这个阶段的主要目标之一，也是根据需要定义一个重新训练模型的过程。

表 7-2 列出了 GINA 团队研究的分析计划。项目主要实现了三个重要的成果。

表 7-2 GINA 团队研究分析计划

项目分析计划	GINA 案例研究
商业理解	跟踪全球知识增长，确保有效的知识转移，并迅速将其转换为公司资产。执行这三个要素应加快创新
初始假设	不同地理区域的知识转移的增加提高了想法交付的效率
数据准备	5 年的创新思想提交和历史；6 个月来自全球创新和研究活动的文字笔记
模型规划分析方法	社交网络分析，社交图，聚类和回归分析
知识发现	(1) 确定隐藏的、高价值的创新者，并找出知识共享的方法 (2) 大学研究项目的投资决策 (3) 创建工具，帮助提交者用创新思维改进推荐系统

创新是每个公司都希望推广的想法，但是很难衡量创新或确定增加创新的方法。这个项目从评估的角度探讨了这个问题，以识别创新网络中的创新卓越者和有影响力的人才。

这个阶段是非常重要的，一方面要检验是否符合预期并能在此基础上提出更合理和更高的目标；另一方面，要检验整个流程过程是否完善，是否有更理想的数据和更好的分析方

法来完成目标。

7.3.4　数据分析工具

在大数据分析领域主要包括模式识别、数据挖掘、预测性分析和可视化分析 4 大类，见表 7-3。

表 7-3　数据分析工具比较

主导优势	操　作		编　程			
	Eviews	SPSS	SAS	Stata	MATLAB	R
	时间序列分析	多元横截面数据	数据管理及挖掘	面板数据处理	数值分析，复杂模型	算法及绘图
应用领域	经济	通信，政府，金融，制造，医药，教育等	市场调研，医药研发，能源公共事业，金融管理等	经济	建筑工程	学术研究，医药研发，IT
处理功能	推断统计	推断及多元统计	批量数据集		统计预测，优化建模	统计分析，数据挖掘
界面设计	直观，可视化	简易，可视化	语言机械规范化	可视，代码灵活	偏向底层	语言丰富灵活
数据安全	软件稳定	大数据易丢失	软件稳定	软件稳定	软件稳定	软件稳定
处理效率	高，稳定	低，不适宜大数据	高，稳定	高，稳定	高，稳定	极适合大量数据
结合形式	Excel，SAS，SPSS	Excel	Excel，文本	文本	所有	所有

1. SAS

SAS 全称为 Statistics Analysis System，最早由北卡罗来纳大学的两位生物统计学研究生编制，并于 1976 年成立了 SAS 软件研究所，正式推出了 SAS 软件。SAS 是用于决策支持的大型集成信息系统，是由大型计算机系统发展而来，其核心操作方式就是程序驱动，经过多年的发展，现在已成为一套完整的计算机语言，其用户界面也充分体现了这一特点。它采用 MDI(多文档界面)，用户在 PGM 视窗中输入程序，分析结果以文本的形式在 OUTPUT 视窗中输出。使用程序方式，用户可以完成所有需要做的工作，包括统计分析、预测、建模和模拟抽样等。但是，这使得初学者在使用 SAS 时必须要学习 SAS 语言，入门比较困难。

目前，SAS 已在全球一百多个国家和地区拥有两万九千多个客户群，直接用户超过 300 万人。在我国，国家信息中心、国家统计局、卫生部，中国科学院等都是 SAS 的大用户。SAS 已被广泛应用于政府行政管理、科研、教育、生产和金融等不同领域，并且发挥着愈来愈重要的作用。

2. 数据挖掘分析工具

Clementine 是 ISL(Integral Solutions Limited)公司开发的数据挖掘工具平台。1999 年，SPSS 公司收购了 ISL 公司，对 Clementine 产品进行重新整合和开发，现在 Clementine 已经成为 SPSS 公司的又一亮点。

Clementine 的图形化操作界面如图 7-7 所示，它使得分析人员能够可视化数据挖掘过程的每一步。通过与数据流的交互，分析人员和业务人员可以合作，将业务知识融入到数据挖掘过程中。这样数据挖掘人员就可以把注意力集中于知识发现，而不是陷入技术任务，例如写代码，所以他们可以尝试更多的分析思路，更深入地探索数据，揭示更多的隐含关系。

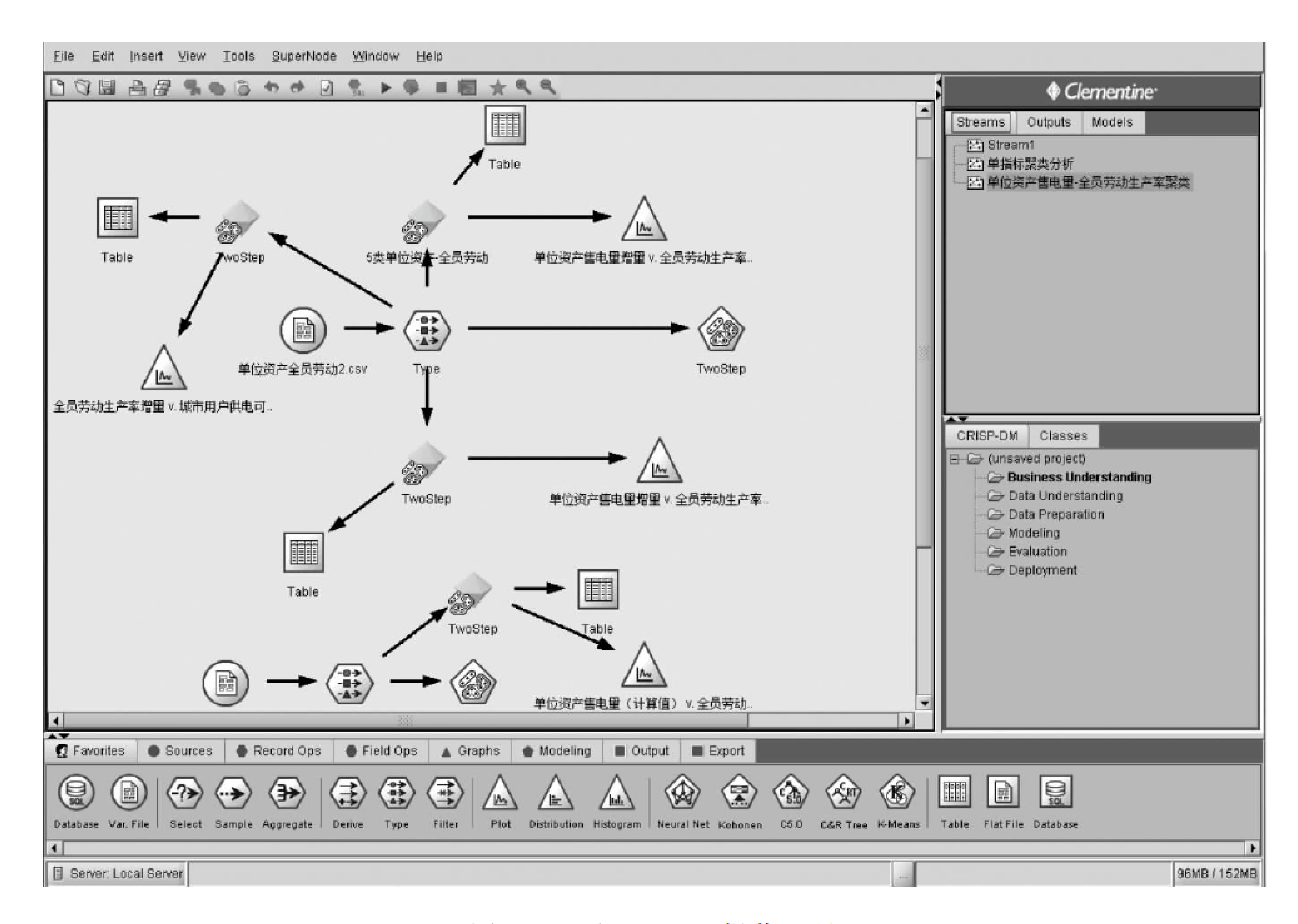


图 7-7 Clementine 操作工具

使用全面整合到 Clementine 的 Text Mining，可以从任何类型的文本——例如，内部报告、呼叫中心记录、客户的邮件、媒体或者杂志文章、博客等中抽取内容和评论。使用 WebMining for Clementine，可以发现访问者网上行为模式。直接获取 Dimension 产品的调查数据，可以把人口统计信息、态度和行为信息用于模型——更深入地理解客户。Clementine 还提供大量的应用模板，例如：

(1) CRM CAT——针对客户的获取和增长，提高反馈率并减少客户流失；

(2) Web CAT——单击顺序分析和访问行为分析；

(3) cTelco CAT——客户保持和增加交叉销售；

(4) Crime CAT——犯罪分析及其特征描述，确定事故高发区，联合研究相关犯罪行为；

(5) Fraud CAT——发现金融交易和索赔中的欺诈和异常行为；

(6) Microarray CAT——研究和疾病相关的基因序列并找到治愈手段。

3. R 语言工具

R 语言是一种自由软件编程语言与操作环境，主要用于统计分析、绘图、数据挖掘。R 本来是由来自新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentlema 开发的，主要是以命令行操作，同时有人开发了几种图形用户界面。现在由“R 开发核心团队”负责开发。R 语言

分析工具如图 7-8 所示。

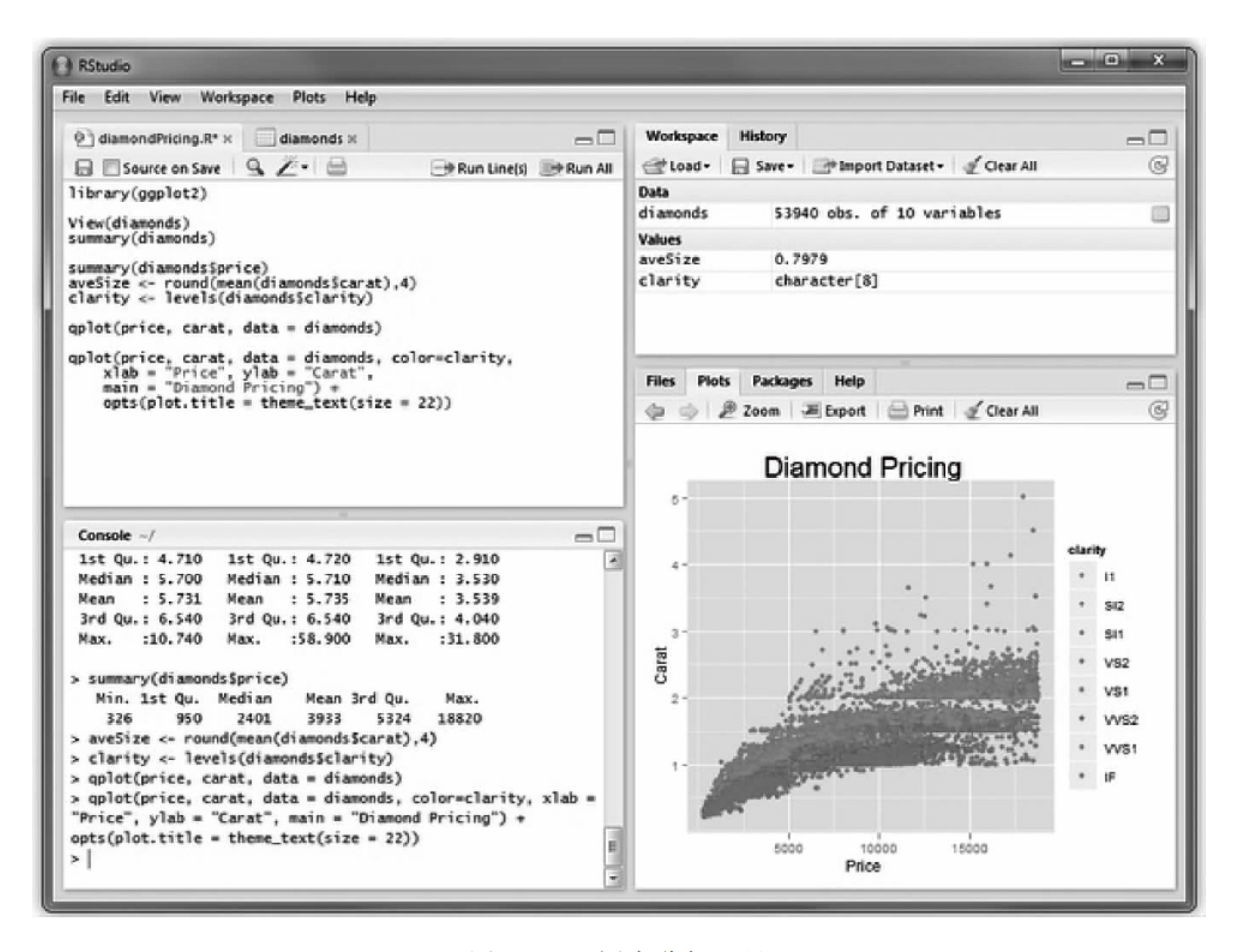


图 7-8 R 语言分析工具

(1) R 内置多种统计学及数字分析功能。R 的功能也可以通过安装包增强。因为具有 S 的血缘,R 比其他统计学或数学专用的编程语言有更强的面向对象(面向对象程序设计)功能。

(2) R 的另一强项是绘图功能,制图具有印刷的素质,也可加入数学符号。

(3) 虽然 R 主要用于统计分析或者开发统计相关的软件,但也有人将其用于矩阵计算。其分析速度可媲美专用于矩阵计算的自由软件 GNU Octave 和商业软件 MATLAB。

4. Stata

Stata 是 Statacorp 于 1985 年开发出来的统计程序,在全球范围内被广泛应用于企业和学术机构中。许多使用者工作在研究领域,特别是在经济学、社会学、政治学及流行病学领域。

作为一个小型的统计软件,其统计分析能力远远超过了 SPSS,在许多方面也超过了 SAS。由于 Stata 在分析时是将数据全部读入内存,在计算全部完成后才和磁盘交换数据,因此计算速度极快(一般来说,SAS 的运算速度要比 SPSS 至少快一个数量级,而 Stata 的某些模块和执行同样功能的 SAS 模块比,其速度又比 SAS 快将近一个数量级)。Stata 也是采用命令行方式来操作,但使用上远比 SAS 简单。其生存数据分析、纵向数据(重复测量数据)分析等模块的功能甚至超过了 SAS。用 Stata 绘制的统计图形相当精美,很有特色。在长远趋势上,Stata 有超越 SAS 的可能。

Stata 最大的缺点应该是数据接口太简单,实际上只能读入文本格式的数据文件;其数据管理界面也过于单调。

5. MATLAB

MATLAB(矩阵实验室)是 MATrix LABoratory 的缩写,是一款由美国 The MathWorks 公司出品的商业数学软件。MATLAB 是一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境。除了矩阵运算、绘制函数/数据图像等常用功能外,MATLAB 还可以用来创建用户界面及调用其他语言(包括 C,C++ 和 FORTRAN)编写的程序,如图 7-9 所示。

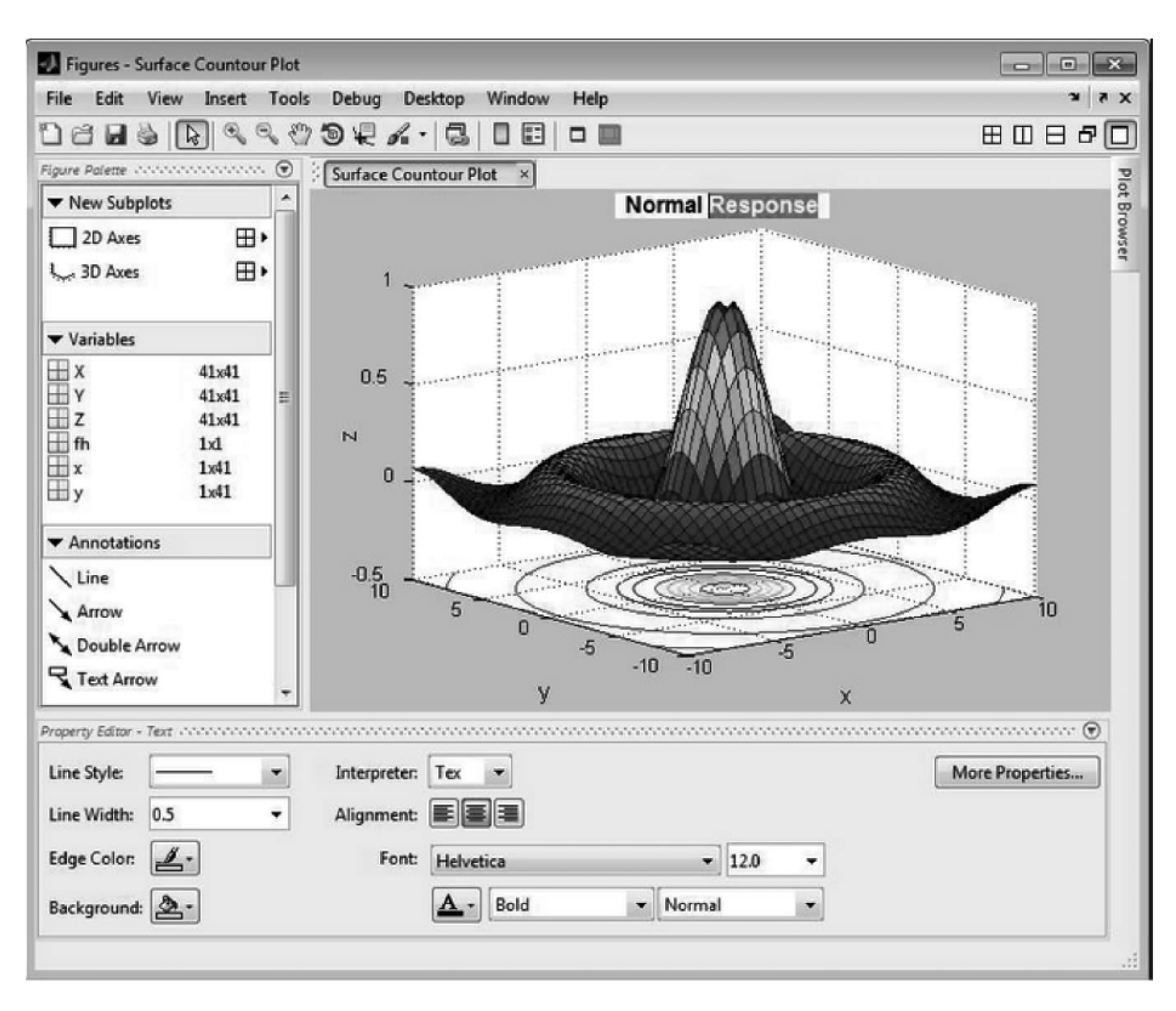


图 7-9 MATLAB 分析工具

MATLAB 和 Mathematica、Maple 并称为三大数学软件。它在数学类科技应用软件中在数值计算方面首屈一指,主要应用于工程计算、控制设计、信号处理与通信、图像处理、信号检测、金融建模设计与分析等领域。

软件特点如下。

(1) 高效的数值计算及符号计算功能,能使用户从繁杂的数学运算分析中解脱出来;

(2) 具有完备的图形处理功能,实现计算结果和编程的可视化;

(3) 友好的用户界面及接近数学表达式的自然化语言,使学习者易于学习和掌握;

(4) 功能丰富的应用工具箱(如信号处理工具箱、通信工具箱等),为用户提供了大量方便实用的处理工具。

7.4 数据仓库理论

数据仓库也是一种数据库，因此传统数据库的原理，比如数据独立性、数据安全性和完整性、并发控制技术等都是数据仓库原理的一部分。本节主要介绍数据仓库本身所有的特征、数据仓库模型、数据仓库设计、数据仓库建设方法论和数据仓库管理相关技术等。

7.4.1 数据仓库的主要特征

根据 William H. Inmon 给出的定义，数据仓库是一个面向主题的、集成的、时变的和非易失的数据集合，支持管理部门的决策过程。根据此定义，数据仓库具有以下 4 个主要特征。

1. 面向主题

数据仓库围绕一些主题如客户、供应商、产品和销售来组织。数据仓库关注决策者的数据建模与分析，而不是组织机构的日常操作和事务处理。

2. 集成

通常，构建数据仓库是将多个异构数据源，如关系数据库、一般文件和联机事务记录集成在一起。使用数据清理和数据集成技术确保命名约定、编码结构及属性度量等的一致性。

3. 时变

数据存储从历史的角度(例如过去 5～10 年)提供信息。数据仓库的关键结构都隐式或显式地包含时间元素。

4. 非易失

数据仓库总是物理地分别存放数据，这些数据源于操作环境下的应用数据，由于这种分离，数据仓库不需要事务处理、恢复和并发控制机制。

总之，数据仓库是语义上一致的数据存储，它充当用于决策支持的数据模型的物理实现，并存放企业战略决策所需要的信息。数据仓库也常常被看作一种体系结构，通过将异构数据源中的数据集成在一起而构造，支持查询、分析报告和决策制定。

7.4.2 数据仓库建模

数据建模是抽象描述现实世界的一种工具和方法，是通过抽象的实体及实体之间联系的形式，来表示现实世界中事务的相互关系的一种映射。数据仓库建模按照应用层次可分为：业务建模、领域建模、逻辑建模和物理建模。

1. 业务建模——主要解决业务层面的分解和程序化

(1) 划分整个单位的业务，一般按照业务部门的划分，进行各个部门之间业务工作的界定，理清各业务部门之间的关系。

(2) 深入了解各个业务部门内具体业务流程并将其程序化。

(3) 提出修改和改进业务部门工作流程的方法并程序化。

(4) 数据建模的范围界定，整个数据仓库项目的目标和阶段划分。

2. 建模领域——主要是对业务模型进行抽象处理，生成领域概念模型

(1) 抽取关键业务概念，并将之抽象化。

（2）将业务概念分组，按照业务主线聚合类似的分组概念。

（3）细化分组概念，理清分组概念内的业务流程并抽象化。

（4）理清分组概念之间的关联，形成完整的领域概念模型。

3. 逻辑建模——主要将领域模型概念实体以及实体之间的关系进行数据层次的逻辑化

（1）业务概念实体化，并考虑其具体的属性。

（2）事件实体化，并考虑其属性内容。

（3）说明实体化，并考虑其属性内容。

4. 物理建模——主要解决逻辑模型针对不同关系型数据库的物理化及性能等技术问题

（1）针对特定物理化平台，做出相应的技术调整。

（2）针对模型的性能考虑，对特定平台做出相应的调整。

（3）针对管理的需要，结合特定的平台做出相应的调整。

（4）生成最后的执行脚本，并迭代完善。

7.4.3 数据仓库设计

数据仓库模型是数据仓库建库和管理，定义数据转移规则和流程，以及设计数据仓库和前端应用接口的重要依据。当数据仓库系统结构需要进行更改时，先检查响应的数据模型，全面了解改动对现有数据仓库结构的影响，然后决定是否需要变化以及怎样变化，再对数据仓库系统的其他模块进行修改。

数据仓库模型的设计一般依据现有主题分析需要，满足需求涉及的数据范畴，从而确定数据集市模型和数据仓库逻辑模型，如图 7-10 所示，主要包括以下几个部分。

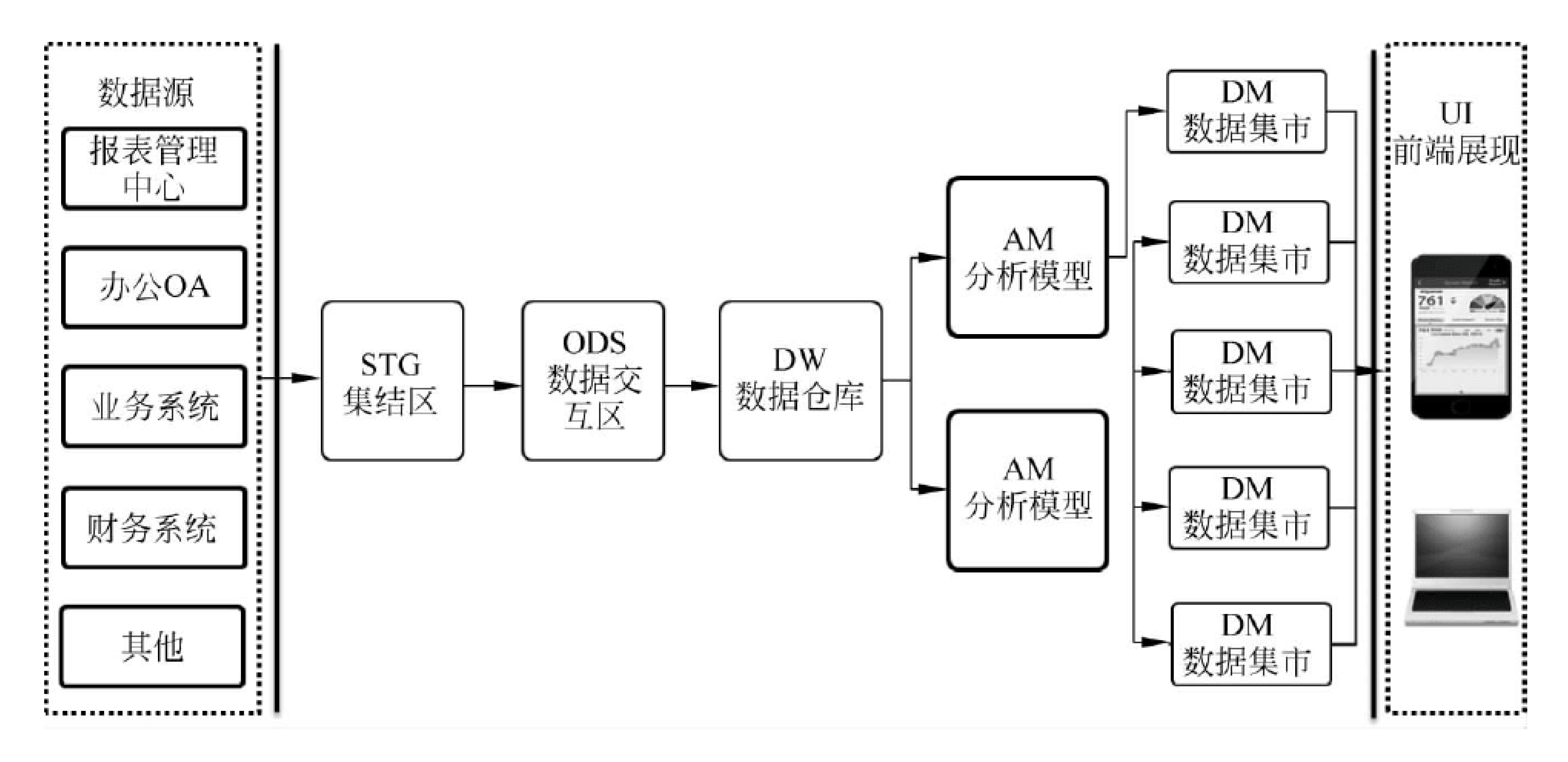


图 7-10　数据仓库模型

（1）数据源：数据采集来源。主要来自报表管理中心、账务系统、其他系统或者文件。

（2）STG 集结区：来源于各个数据源的数据存放区域，只存储最近变化的一个时间段内的数据、未经处理和整合的数据。

（3）ODS 交互区：来源于 STG 集结区，将集结区内的数据转化为通用的格式在 ODS 内进行存储，将来源于各个业务系统的数据按照通用的格式进行数据处理和转换。

(4) DW 数据仓库：数据结构与 ODS 交互区的数据结构一致，存储长期稳定的数据，最新的数据根据来自 ODS 交互区的数据进行更新。

(5) AM 分析模型：来源于 DM 数据仓库，将 DW 数据仓库的数据进行转换、加工处理，作为数据集成的基础。

(6) DM 数据集市：根据不同的业务主题，将 AM 分析模型的数据按照不同的主题分别存储到不同的数据集市中，供前端展示使用。

(7) UI 前端展现：语义层和前端展现。

7.4.4　数据仓库建设方法论

数据仓库的建设是一个复杂的系统工程。从数据仓库技术诞生到现在，有许多企业进行了数据仓库的建设，有很多成功的经验，也有很多失败的教训。企业如何有效地建立数据仓库需要具体方法论指导。业界有很多厂商和公司都在其数据仓库的建设中积累了很多的经验知识，并形成了一些系统的方法论。

Oracle 数据仓库建设方法论(DWM)是一个结构化的实施方法，定义了用于构件一个完善的、满足业务功能的数据仓库系统所需要的典型步骤和任务。Oracle 数据仓库建设方法论中的增量实施法把数据仓库系统的实施分为 13 个过程，7 个阶段，示例如表 7-4 所示。

表 7-4　数据仓库方法论

过 程 阶 段	实施策略	系统定义	系统分析	系统设计	系统建立	系统应用	系统维护
业务需求定义	34.6%	9.3%	13.3%				
数据获取	8.4%	8.5%	23.1%	16.4%	17.2%	21.4%	
系统结构定义	11.5%	22.2%	14.3%	5.4%	14.5%		
数据质量控制	2.2%	12.1%	6.5%	7.9%	0.6%		
数据仓库管理	3.4%	4.4%	3.9%	11%	16.1%		
元数据管理	3.7%			4.7%	4.9%		
数据访问	6.3%			4.7%	4.9%		
数据库设计与建立				4%	2%		
文档设置	1%	1.1%	2%	4.5%	3.9%		
系统测试	1.2%		7.4%	15%	19.8%	19.7%	
培训	0.5%	7.7%	2.7%	4.1%	4.7%	27.8%	
系统上线			1.7%	0.2%	0.4%	17.8%	
技术支持							45%

表 7-4 中横向为数据仓库建设的 7 个阶段，纵向为 13 个建设过程。从策略规划到最后的系统维护阶段，涵盖了数据仓库项目建设的全生命周期阶段。由于数据仓库的建设活动过程基本上都是跨阶段实施完成的，所以中间的纵横交叉点则是实施过程在每个实施阶段的分布，各个过程右边的区域表示每个过程涉及哪个实施阶段，表中的百分比表示以一个中等规模的项目为例，每个过程在各个实施阶段中所占的比例。DWM 可以帮助我们解决诸如确定正确的系统范围和用户需求，建立灵活的系统架构以满足不断变化的应用需求和不可预测的使用需求等数据仓库建设中的问题。

NCR Teradata 数据仓库建设方法论是一个系统的体系。该方法论使整个数据仓库的

实施处于完全可控制状态，方法论描述了实施的各个步骤。方法论包括4个阶段：数据仓库策略开发、数据仓库规划、数据仓库设计和实现、数据仓库支持和增强，如图7-11所示。

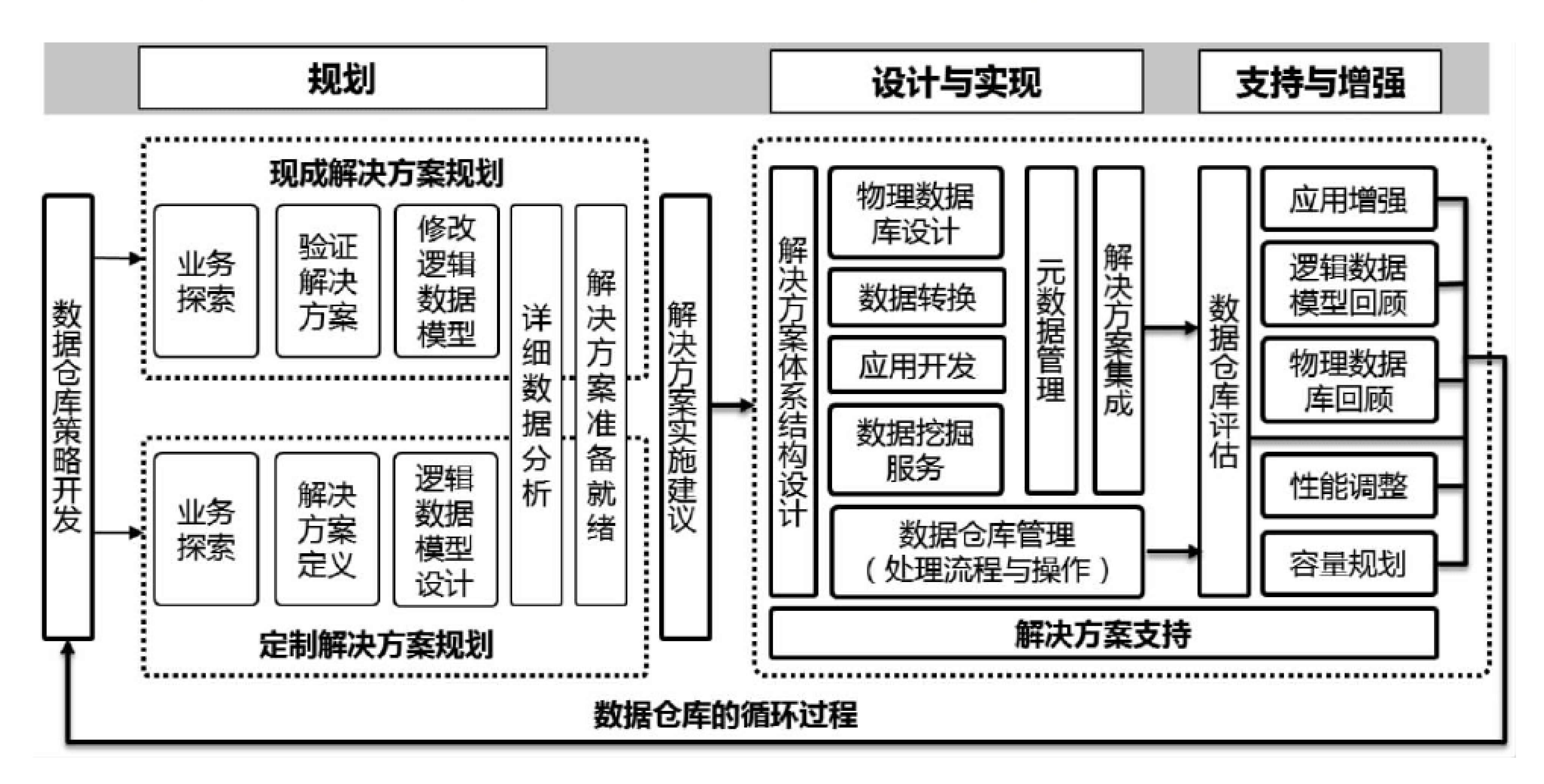


图7-11　NCR数据仓库建设方法论

数据仓库策略开发是数据仓库建设方法论的起点，构造了后续数据仓库活动的道路。接下来，规划阶段执行对特定业务领域的详细的分析和设计。分析和设计完成后，设计与实现阶段建造出有业务价值的实际数据仓库。在数据仓库投入运行后，开始进入持续的开发维护阶段。NCR数据仓库建设方法论是一个循环的过程，以实现数据仓库和业务的持续改进。

7.4.5　数据仓库相关技术

1. 数据提取、转换与加载

数据仓库系统是在业务系统的基础上发展起来的，其内部存储的数据来自于事务处理的业务系统和外部数据源。因企业的业务系统是在不同时期、不同背景、面对不同应用、不同开发商等各种客观前提下建立的，其数据结构、存储平台、系统平台均存在很大的异构型。这导致各种数据源缺少统一标准，因而其数据难以转化为有用的信息，原始数据的不一致导致决策时其可信度降低。

此外，随着企业的不断发展，既有的业务系统、业务流程以及相关的信息结构都可能会发生变化，这种变化将直接影响到后端数据仓库系统中的数据更新。如何有效地维护这种变化，尽量控制数据仓库刷新操作的成本，也是数据仓库建构中极为重要的一个问题。

ETL是建构企业数据仓库从而实现商业智能的核心和灵魂，它按照统一的规则集成数据并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。

ETL的基本功能如下。

（1）数据提取：全量提取、增量提取（触发器、时间戳、全表比对、日志解析）。

（2）数据转换：工具引擎内转换和通过SQL转换。

(3) 数据加载：直接使用 SQL 语句进行 DML 操作和采用批量装载方法。

2. OLAP 多维分析

从 20 世纪 80 年代开始，被称为联机事务处理(On-Line Transaction Processing，OLTP)的数据库应用系统已经在企事业单位得到广泛应用。为了充分利用 OLTP 数据库中大量的数据，并为企业提供更加准确且多角度的决策信息，关系数据库之父 E. F. Codd 及其同事于 1993 年提出了联机分析处理(On-Line Analysis Processing，OLAP)的概念。OLAP 是针对特定问题的联机多维数据快速访问和分析处理的软件技术，帮助决策者方便地对数据进行深入的多角度观察和分析利用。

(1) OLAP 基本功能。OLAP 是用户进行决策分析最重要的工具，通过对多维数据采用切片、切块、上钻、下钻、旋转等分析操作，从多个角度观察经营数据，从而深入地了解包含在数据中的信息和内涵。

(2) 多维数据模型。多维数据模型将数据看作数据立方体形式，允许用户从多维度对数据建模和观察。多维数据模型由维度和度量定义。一般地，维度是一个组织想要保存记录的透视图和实体，每个维度都有一个表与之相关联，称为维表，它是对维的进一步描述。维表可以由用户设定，或者根据数据分布自动产生和调整。通常，多维数据模型围绕中心主题组织，主题用事实表表示。事实表包括度量和每个相关维表的键。数据仓库需要简明的、面向主题的模型，便于联机数据分析。最流行的数据仓库数据模型是多维数据模型。它一般分为星型模型、雪花模型等。

① 星型模型。星型模型是最常用来表示多维数据的一种模型。在该模型中，一个多维数据模型包含一张事实表和多维表，每个维对应于一张维表。事实表中的每个元组包含一个度量值和一组指针，有多少个维度就有多少个指针，这些指针分别指向相应维表中对应于该元组的那条记录，这种指针在关系模型中通常使用外键来表示，维表由描述这个维度的各个属性组成。图 7-12 给出了星型模型的一个实例。

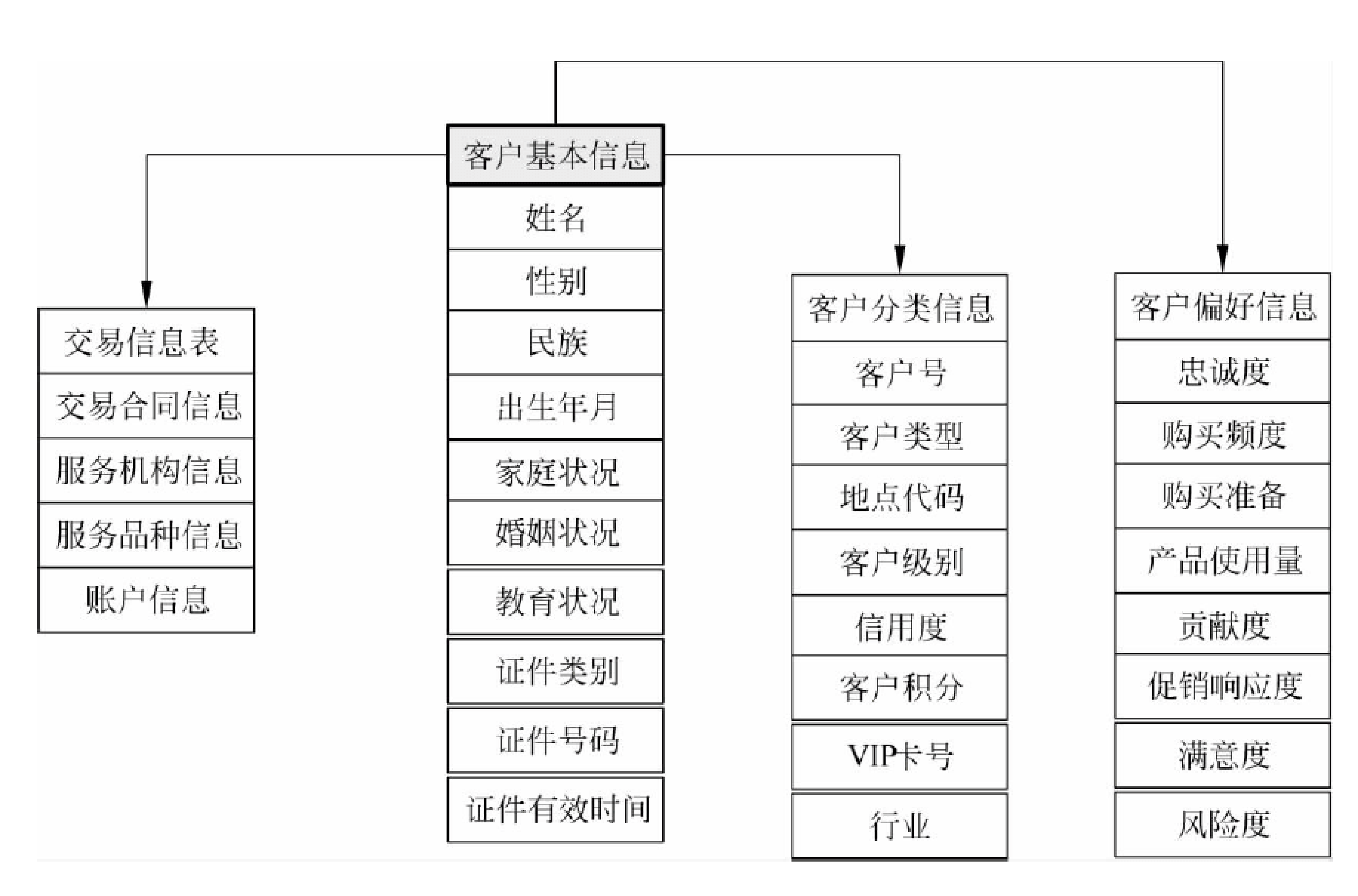


图 7-12　星型模型实例

事实表一般行数较多,而维度表相对来说行数较少。星型模型存取数据速度快,针对各个维做了大量的预处理,如按照维度进行预先的统计、分类和排序等。

② 雪花模型。雪花模型是对星型模型的扩展,雪花模型对星型模型的维表进一步层次化,原来的各维表可能被扩展为小的事实表,形成一些局部的"层次"区域。它的优点是最大限度地减少数据存储量,以及把较小的维表联合在一起来改善查询性能。

雪花模型增加了用户必须处理的表的数量和某些查询的复杂性,但这种方式可以使系统更进一步专业化和实用化,同时也降低了系统的通用程度。前端工具将用户的需求转化为雪花模型的物理模式,完成对数据的查询。

如图 7-13 所示,在星型模型的基础上,对"产品维度表"进行扩展,形成雪花模型。既满足了用户对复杂数据仓库查询的需求,又能够完成一些简单查询功能而不用访问过多的数据。

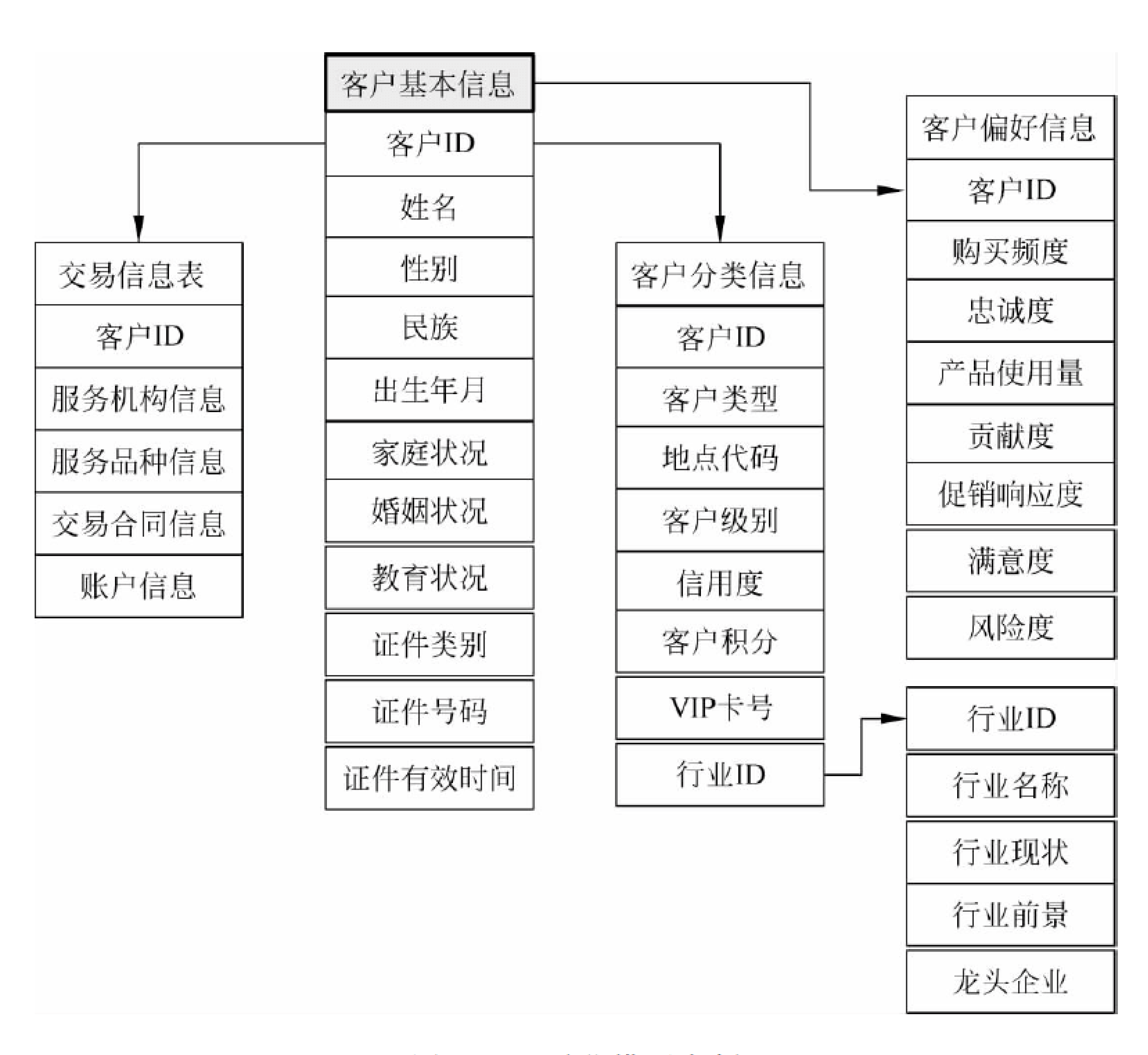


图 7-13　雪花模型实例

7.4.6　DW、OLAP 与 DM 的关系

DW(数据仓库)、OLAP(联机分析处理)与 DM(数据挖掘)是相互独立而又相关联系的三个概念。相互独立指它们是在不同的时期产生的,由不同的学者或机构分别提出,因此,它们在概念内涵上、主要解决问题上以及使用技术上都有很大差别。相互联系是因为它们都是为了支持企业的管理决策而提出的。它们三者的关系见表 7-5。

表 7-5　DW、OLAP 与 DM 的关系

主要不同点	数据仓库(DW)	OLAP	数据挖掘(DM)
提出的时间	1991 年	1993 年	1989 年
提出的学者	W. H. Inmon(恩门)	E. F. Codd	第 11 届国际人工智能联合会
概念的内涵	数据集成的历史数据集合	历史数据的多维分析和展示方法	挖掘数据中隐藏知识的算法或工具
解决的问题	海量数据的集成、组织和存储	人机交互数据的多维联机分析处理	数据中隐藏知识的发现问题
主要的技术	数据库及相关技术	数据工程与统计分析技术	机器学习、模式识别等人工智能技术

7.5　数据挖掘高级理论

在了解数据挖掘的概念时，首先要知道知识发现(Knowledge Discovery in Database，KDD)的定义。知识发现就是采用有效的算法从大量的、不完全的、有噪声的、模糊和随机的数据中识别出有效的、潜在价值的并能形成最终可理解的模式(Pattern)的非平凡过程。

知识发现的过程一般包括数据采集、数据选择、数据整合、数据挖掘、知识评价和知识应用等主要步骤。

数据挖掘是知识发现过程中的一个重要而关键的步骤。但现在的文献大多对这两个术语不加区分地使用，并且在大多数场合都用数据挖掘术语代替知识发现。本书后面章节也将数据挖掘等同于知识发现。此外，基于大数据的数据挖掘对象包括结构化数据和非结构化数据。基于大数据的数据挖掘分析方法主要用于预测未来。预测未来的分析方法主要包括聚类分析、分类分析、关联分析、时序模型、结构优化和机器学习等。

7.5.1　聚类分析

聚类分析(Clustering Analysis)，也称为群集分析，是用于静态数据分析的一门技术。聚类是把相似的对象通过静态分类的方法分成不同的组别或更多的子集，这样在同一个子集的成员对象都有相似的一些属性。常见的有各种各样的距离的算法，但基于距离的算法的一个致命的缺点就是只能发现“类圆形”的聚类，因此，后来又有人提出了机遇密度的聚类。一般把数据聚类归纳为一种非监督式学习，它是指在没有分类标签的数据中寻找内在关联。聚类分析是一门交叉学科，它被广泛应用在统计学、机器学习和数据挖掘等相关领域之中，特别是在数据挖掘领域，吸引了很多的研究者进行此相关课题的研究。

到目前为止，国内外的学者提出许多关于聚类的分析方法，但是，整体来讲，聚类的方法可以被区分为：划分方法，层次方法，密度方法，网格方法和模型方法等。

聚类算法通常定义为：假设有一组数据集或者大量的数据，怎样通过一种无监督的方法，把数据集或者大量数据进行不同的区分，也就是说，根据一定的衡量方式，把相似性很高的数据和相似性较低的数据划分为不同的类。在数据的分析过程中，根据语义的不同，数据分析可以分为：聚类类簇和聚类分析。它们的主要区别是：类簇是类别，而分析是技术方法。在当前，聚类分析研究被广泛应用到了许多领域当中，因而，在解决许多不同类型的问题时，衍生出了许多不同的聚类方法，这些方法都有不同的特性；因此，当在实际生活当中

应用到不同聚类算法时,需要结合实际,考虑多方面的因素来选取聚类算法,例如,数据的规模、领域、效率等条件,从而为做出决策提供关键性的数据支撑。

1. *k*-means 聚类

k-means 是聚类分析的经典算法之一,主要是作为一种探索式的技术,用来发现之前没有被注意到的数据结构。尽管在聚类中记录的类别不是已知的,但是聚类可以用来探索数据的结构,总结类群的属性特征。当维度比较低的时候,我们可以可视化类群(Cluster),但随着维度增加,可视化类群就越来越困难。*k*-means 聚类有很多应用,包括模式识别、人工智能、图像处理、机器视觉等。

假设输入数据由多条记录组成,每条记录包含多个数字,即每条记录可以看成一个由数字组成的向量。要衡量两个向量之间的相似度,就要选择一种度量方式。选择度量时通常有以下几个原则。

(1) 距离为非负;

(2) 同一个向量之间的距离为零;

(3) 向量 p 到向量 q 之间的距离与向量 q 到向量 p 之间的距离相等;

(4) 三个向量之间的距离,任意一个距离不大于另两者之和。

欧几里得距离是一种最流行的距离度量方法。向量 p 和向量 q 之间的距离:

$$d(p,q)=d(q,p)=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

从以上表达式可以看出:

第一,欧几里得距离受变量的规模影响,改变变量的规模(比如从厘米到毫米),可以显著地影响结果。

第二,欧几里得距离不考虑变量之间的关联。

第三,算法对离群值很敏感,意味着如果数据中有离群值且无法去掉的话,聚类的结果会受严重的影响。

下面来看 *k*-means 聚类算法过程,如图 7-14 所示。

第一步:随机选择 K 个"中心点"。

第二步:将每条记录分配到最近的"中心点"上,形成类群(Cluster)。

第三步:在第二步基础上重新计算新的类群的中心点,中心点的属性为类群中记录的均值。

第四步:重复第二步和第三步直至中心点不再改变。

从 *k*-means 算法中可以看出,选择一个正确且合适的 K 值,有助于算法正确地将记录分类。或者,可以重复地尝试不同的 K 值,从中选出一个最佳的。当然,一些专业领域的知识可以帮助确定 K 的值。

当数据之间的聚类不明显,以至于 K 值很难确定时,可以使用一种启发式的方法来挑选最优的 K 值和组内的平方和(Within Sum of Squares,WSS),WSS 是一种用来衡量聚类之间有多紧密的方式,即残差的方差,公式如下:

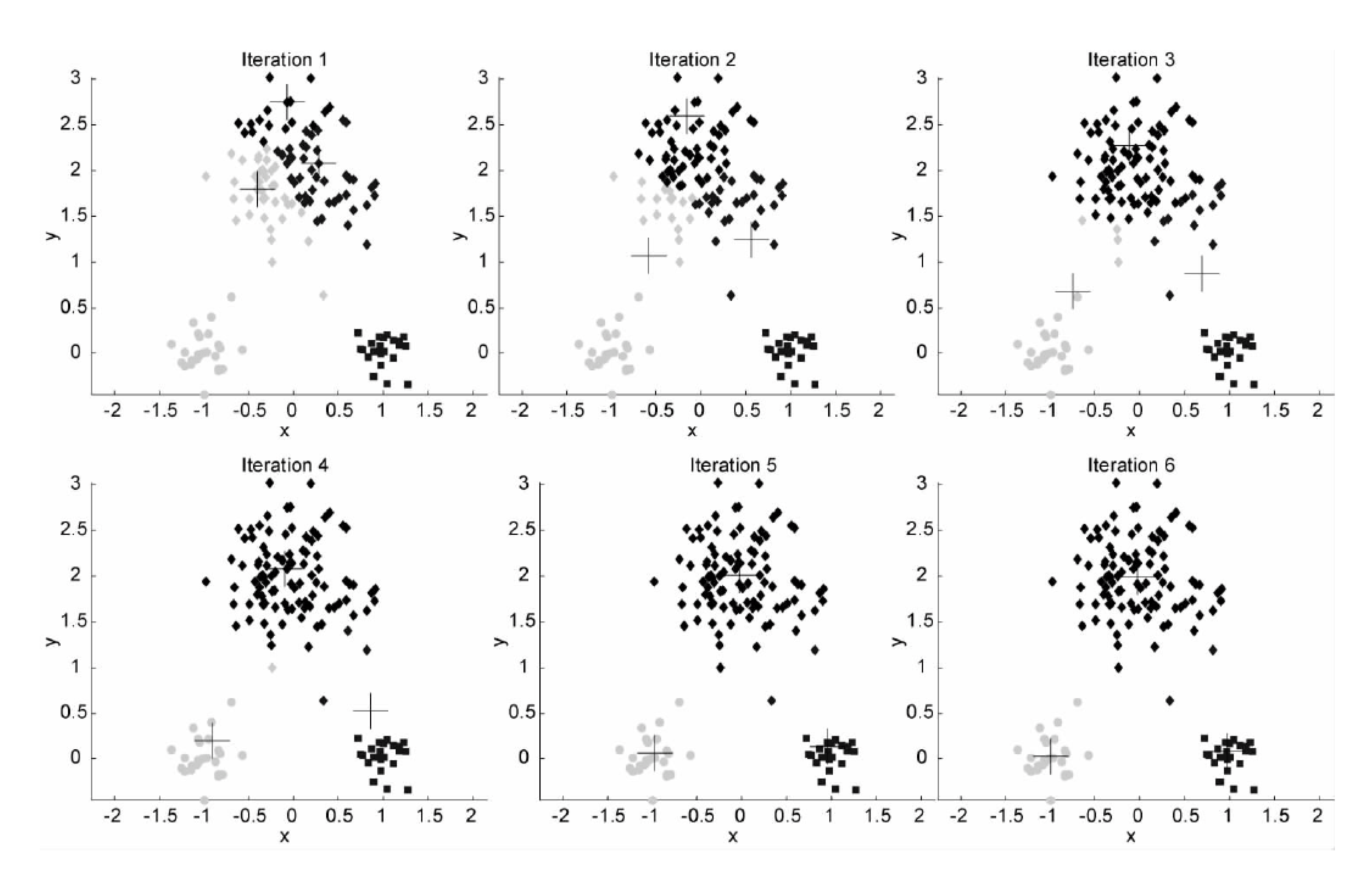


图 7-14　*k*-means 聚类算法过程示例

$$\mathrm{WSS}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}|\chi_{ij}-c_i|^2$$

WSS 即是计算每个聚类中每个点与中心点的差的绝对值的平方和的求和。式子中的 χ_{ij} 是类中的点，c_i 是类的中心。通常更多的类（即更大的 K 值），会使得每个类更“紧密”，但类太多会带来过拟合问题。意味着 WSS 值会随着 K 值的增大而减小，但有时候会上升，在上升前的拐点是个比较好的选择。

2. MapReduce 形式的 *k*-means 聚类

利用 MapReduce 计算模型，可以把 *k*-means 应用到大数据中进行数据挖掘。MapReduce 形式的 *k*-means 也很简单，每执行一次 MapReduce 作业的时候，重新迭代计算中心点，直到中心点不再改变为止。

先随机产生一组中心点，然后在开始执行的时候加载。

Map 阶段：

(1) 载入中心点。

(2) 计算每行数据与中心点的距离。

(3) 为每行数据挑选一个距离最近的中心点。

(4) 输出：Key＝中心点；Value＝本行数据。

Reduce 阶段：

(1) 遍历中心点，重新计算中心点位置。

(2) 输出：中心点。

重复以上过程，就可以得到聚类的中心点。当然，K 值也是需要事先指定的。

但是，MapReduce 形式的 k-means 算法也有一些缺点，还存在优化的空间。

（1）每次循环都重复读入相同的数据集；

（2）没有针对算法进行计算本地化优化；

（3）MapReduce 处理循环迭代任务的效率不高。

为了克服这几个缺点，华盛顿大学（University of Washington）的 Bill Howe 教授领导的 HaLooP 项目就针对此类问题进行了优化。HaLoop 对 Hadoop 中的 MapReduce 框架进行了修改，使其能够适应迭代循环的任务，同时保留了 MapReduce 框架容错的特性。HaLoop 架构如图 7-15 所示。

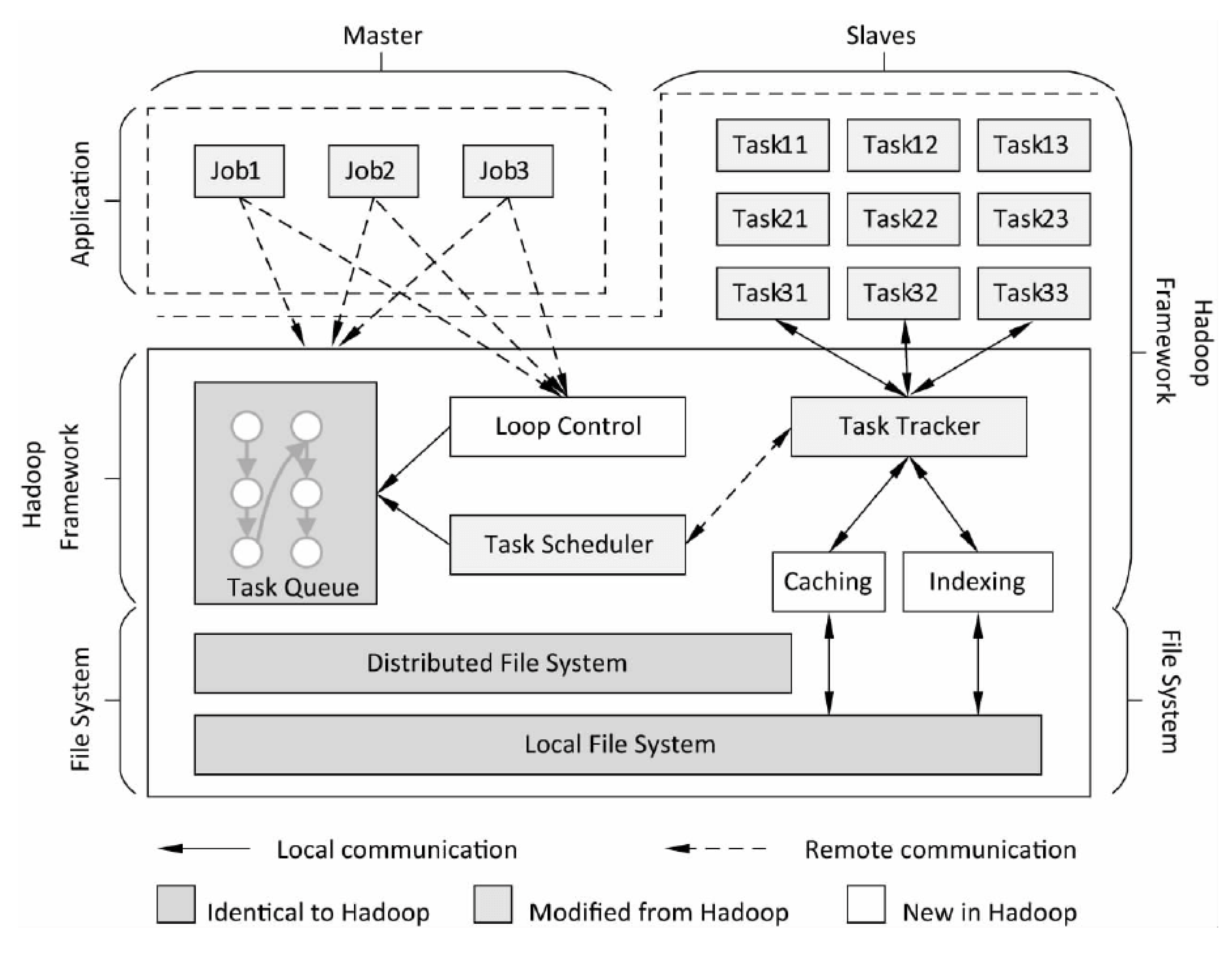


图 7-15　HaLoop 架构图

将 HaLoop 应用在 k-means 算法上，优势在于 Mapper 输入时的缓存机制。HaLoop 的 Mapper 输入缓存目的在于避免非本地的数据被读入到 Mapper 中（第一次循环除外）。在第一次循环时，如果 Mapper 读入了非本地数据，那么 Mapper 会将其加入缓冲中，然后在接下来的循环中读取。下面是利用 HaLoop 来优化 k-means 算法的示例。

```
Public class KMeansLoopInputOutput implements LoopInputOutput{
    @Override
    Public List<String> getInputPaths(JobConf conf, int iteration, int step){
        List<String> paths = new ArrayList<String>();
        //only input the dataset,cluster means are
        //read from HDFS in Mappers
        Paths.add(conf.getInputPath());
        Return paths;
```

```
    }
@Override
   Public String getOutputPath(JobConf conf,int iteration,int step){
      Return (conf.getOutputPath() + "/i" + iteration);
   }
}
Public class KMeansLoopMapCacheFilter implements CacheFilter{
@Override
       Public Boolean isCache(Object key,Object value,int id){
       //cache every tuple
           Return true;
        }
}
```

本节介绍了 k-means 聚类在数据挖掘中的应用，以及如何利用分布式计算来进行 k-means 聚类。现在来总结一下 k-means 聚类的优缺点。

优点：

(1) 实现简单；

(2) 容易将新的数据分配给已有的类别，寻找最近的聚类即可；

(3) 输出简洁，只有 K 个中心点。

缺点：

(1) 不能处理分类变量；

(2) 对第一次中心点的分步敏感；

(3) 变量都应该以相同或相似的方法来衡量，衡量方法不可变；

(4) K 值必须已知，错误的猜测会导致错误的结果；

(5) 趋向于产生等大小的聚类的结果，不一定足够理想。

7.5.2　关联分析

关联分析又称关联挖掘，就是在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。或者说，关联分析是发现交易数据库中不同商品(项)之间的联系。关联分析是一种简单、实用的分析技术，就是发现存在于大量数据集中的关联性或相关性，从而描述了一个事物中某些属性同时出现的规律和模式。

关联规则是另外一种无监督学习的方法，同样没有“预测”的过程，主要用于发现数据之间的联系。典型的应用场景有：

(1) 哪些商品通常会被一同购买？

(2) 喜欢/购买了这个产品的顾客会倾向于喜欢/购买哪些其他产品？

关联分析的一个典型例子是购物篮分析。该过程通过发现顾客放入其购物篮中的不同商品之间的联系，分析顾客的购买习惯。了解哪些商品频繁地被顾客同时购买，通过这种关联的发现可以帮助零售商制定营销策略。其他的应用还包括价目表设计、商品促销、商品的排放和基于购买模式的顾客划分。

可从数据库中关联分析出形如"由于某些事件的发生而引起另外一些事件的发生"之类的规则。如"67%的顾客在购买啤酒的同时也会购买尿布"，因此通过合理的啤酒和尿布的货架摆放或捆绑销售可提高超市的服务质量和效益。又如"C 语言课程优秀的同学，在学习'数据结构'时为优秀的可能性达 88%"，那么就可以通过强化"C 语言"的学习来提高教学效果。

关联规则挖掘的目标是寻找数据之间"有价值"的关联。"有价值"取决于用来挖掘的算法。关联规则的表达形式是，当某人单击/购买产品 X 时，也倾向于单击/购买产品 Y。在这个过程中，有两个关键阈值用来评估关联规则的重要度，即支持度和置信度。

支持度，即项在数据集中的频度： $\text{support}(A)=\frac{\text{support_count}(A)}{|D|}$

置信度： $\text{confidence}(A\Rightarrow B)=\frac{\text{support_count}(A\cup B)}{\text{support_count}(A)}$

其中，support_count(A)指的是 A 在数据集 D 中出现的次数。支持度表达的意思是数据集中 A 出现的频度。置信度表达的意思是在 A 出现的基础上，既出现 A 也出现 B 的频度。

关联规则挖掘通常被运用于交易数据集，由离散的项(Item)组成，比如：

(1) 零售交易数据集；

(2) 一天内计划完成的任务；

(3) 用户的会话中单击的链接组成的日志。

以下主要介绍几种分类挖掘的方法。

1. Apriori 算法

Apriori 算法是挖掘产生布尔关联规则所需频繁项集的基本算法，也是最著名的关联规则挖掘算法之一。Apriori 算法就是根据有关频繁项集特性的先验知识而命名的。它使用一种称作逐层搜索的迭代方法，k-项集用于探索($k+1$)-项集。首先，找出频繁 1-项集的集合，记作 L_1，L_1 用于找出频繁 2-项集的集合 L_2，再用于找出 L_3，如此下去，直到不能找到频繁 k-项集。找每个 L_k 时需要扫描一次数据库。为提高按层次搜索并产生相应频繁项集的处理效率，Apriori 算法利用了一个重要性质，并应用 Apriori 性质来帮助有效缩小频繁项集的搜索空间。

Apriori 性质：一个频繁项集的任一子集也应该是频繁项集。证明根据定义，若一个项集 I 不满足最小支持度阈值 min_sup，则 I 不是频繁的，即 $P(I)<\text{min_sup}$。若增加一个项 A 到项集 I 中，则结果新项集($I\cup A$)也不是频繁的，在整个事务数据库中所出现的次数也不可能多于原项集 I 出现的次数，因此 $P(I\cup A)<\text{min_sup}$，即($I\cup A$)也不是频繁的。这样就可以根据逆反公理很容易地确定 Apriori 性质成立。

针对 Apriori 算法的不足，对其进行优化。

(1) 基于划分的方法。该算法先把数据库从逻辑上分成几个互不相交的块，每次单独考虑一个分块并对它生成所有的频繁项集，然后把产生的频繁项集合并，用来生成所有可能的频繁项集，最后计算这些项集的支持度。这里分块的大小选择要使得每个分块可以被放入主存，每个阶段只需被扫描一次。而算法的正确性是由每一个可能的频繁项集至少在某一个分块中是频繁项集保证的。

上面所讨论的算法是可以高度并行的。可以把每一分块分别分配给某一个处理器生成频繁项集。产生频繁项集的每一个循环结束后.处理器之间进行通信来产生全局的候选是1 项集。通常这里的通信过程是算法执行时间的主要瓶颈。而另一方面,每个独立的处理器生成频繁项集的时间也是一个瓶颈。其他的方法还有在多处理器之间共享一个杂凑树来产生频繁项集,更多关于生成频繁项集的并行化方法可以在其中找到。

(2) 基于 Hash 的方法。Park 等人提出了一个高效地产生频繁项集的基于杂凑(Hash)的算法。通过实验可以发现,寻找频繁项集的主要计算是在生成频繁 2-项集 L_k 上,Park 等就是利用这个性质引入杂凑技术来改进产生频繁 2-项集的方法。

(3) 基于采样的方法。基于前一遍扫描得到的信息,对它详细地做组合分析,可以得到一个改进的算法,其基本思想是:先使用从数据库中抽取出来的采样得到一些在整个数据库中可能成立的规则,然后对数据库的剩余部分验证这个结果。这个算法相当简单并显著地减少了 FO 代价,但是一个很大的缺点就是产生的结果不精确,即存在所谓的数据扭曲(Dataskew)。分布在同一页面上的数据时常是高度相关的,不能表示整个数据库中模式的分布,由此而导致的是采样 5%的交易数据所花费的代价同扫描一遍数据库相近。

(4) 减少交易个数。减少用于未来扫描事务集的大小,基本原理就是当一个事务不包含长度为 L_k 的大项集时,则必然不包含长度为 L_{k+1} 的大项集。从而可以将这些事务删除,在下一遍扫描中就可以减少要进行扫描的事务集的个数。这就是 AprioriTid 的基本思想。

比如,输入一个最小的支持度阈值,只有满足这个阈值的关联规则才会被挖掘出来。Apriori 算法利用的是这样一个特性:任何频繁项集的子集都是频繁的。比如,当我们挖掘出(A,B,C)的支持度满足阈值,即是频繁项集的时候,那么它的任何一个子集,比如(A,B)或(A,C)都是频繁的,因为凡是出现了(A,B,C)的数据中也一定出现了(A,B)或(A,C)。遵循这个思想,Apriori 算法有效地精简了搜索空间。

Apriori 算法的步骤如下。

(1) 在最小的支持度阈值的基础上,找出 1 项的频繁项集,然后找到两个频繁项之间的组合及组合的支持度。

(2) 精简掉所有不符合最小支持度的项集。

(3) 逐步利用频繁项的组合增加项的个数,并重复以上过程,直到找到所有的频繁项集或项集中项的个数达到最大值。

应用案例:信用卡记录数据集的 Apriori 算法挖掘关联规则。

假设:

(1) 1000 条信用记录;

(2) 最小支持度为 0.5,即只有出现频率达到 50%或以上(支持计数达到 500 或以上)的项才会被考虑。

找出一个元素的且符合最小支持度的项集,如表 7-6 所示。

将表 7-6 中支持计数不满足 500 的项集去掉。接下来把这些项组合在一起,然后再遍历数据集得到支持计数,如表 7-7 所示。

表 7-6　寻找 1 项的频繁项集

频繁项集	支持计数
credit_good	700
credit_bad	300
male_single	550
male_mar_or_wid	92
female	310
job_skilled	631
job_unskilled	200
home_owner	710
renter	179

表 7-7　寻找 2 项的频繁项集

频繁项集	支持计数
credit_good,male_single	402
credit__good,job_skilled	544
credit__good,home_owner	527
male_single,job_skilled	340
male_single,home_owner	408
job_skilled,Home_owner	452

去掉不满足最小支持度的频繁项集，得到频繁项集{credit_good,job_skilled}和{credit__good,home_owner}，更进一步得到三个元素的集合，如表 7-8 所示。

表 7-8　寻找 3 项频繁项集

频繁项集	支持计数
credit_good,job_skilled,home_owner	402

由于不满足最小支持度，到这里就没有项集了。

从两个元素的频繁项集中，可以得到以下候选规则，如表 7-9 所示。

(1) credit_good=> job_skilled

(2) job_skilled=> credit_good

(3) credit_good=> home_owner

(4) home_owner=> credit_good

表 7-9　候选规则

规　则	项　集	支持计数	项　集	支持计数	置信度
credit_good=> job_skilled	credit_good	700	credit_good,job_skilled	544	544/700=77%
credit_good=> home_owner	credit_good	700	credit_good,home_owner	527	527/700=75%
job_skilled=> credit_good	job_skilled	631	job_skilled,credit_good	544	544/631=86%
home_owner=> credit_good	home_owner	710	home_owner,credit_good	527	527/710=74%

从表 7-9 中，可以看到 job_skilled=> credit_good 有一个较高的置信度 86%，是比较可靠的规则。

从以上案例中，可以发现 Apriori 算法的优缺点，如表 7-10 所示。

表 7-10 Apriori 算法的优缺点比较

优　　点	缺　　点
实现简单	需要多次遍历数据集
有效地精简了搜索空间(任何频繁项集的自己都是频繁的)	指数的时间复杂度
容易并行化	容易得到伪造或巧合的关联
	可能产生大量候选集

2. FP-growth 算法

由于 Apriori 方法的固有缺陷,即使进行了优化,其效率也仍然不能令人满意。2000 年,Han Jiawei 等人提出了基于频繁模式树(Frequent Pattern Tree,FP-tree)的发现频繁模式的算法 FP-growth。在 FP-growth 算法中,通过两次扫描事务数据库,把每个事务所包含的频繁项目按其支持度降序压缩存储到 FP-tree 中。在以后发现频繁模式的过程中,不需要再扫描事务数据库,而仅在 FP-Tree 中进行查找即可,并通过递归调用 FP-growth 的方法来直接产生频繁模式,因此在整个发现过程中也不需产生候选模式。该算法克服了 Apriori 算法中存在的问题,在执行效率上也明显好于 Apriori 算法。

从上面的对 Apriori 算法的介绍中,可以看出 Apriori 有明显的缺点,就是可能会产生大量的候选集。例如,前一步产生了 10^4 个 1 项的候选集,则会产生 10^7 个 2 项的候选集。另一种不产生候选集的挖掘关联规则的方法就是 FP-growth。

FP-growth 指的是 Frequent-Pattern growth(频繁模式增长),采用的是一种分而治之的方法,通过构建一个紧凑的 FP-tree,然后再 FP-tree 上进行挖掘。

先遍历一次数据集,对数据集中的每项计数。与 Apriori 的 1 项一致,其结果就是每项的支持计数。然后根据最小支持度阈值,将不满足最小支持度阈值的项去掉。

接下来对上一步产生的项进行排序,依据它们的支持度从大到小排序。组成一个 Header Table。

然后再次遍历数据集,构建 FP-tree。FP-tree 的结构如图 7-16 所示。

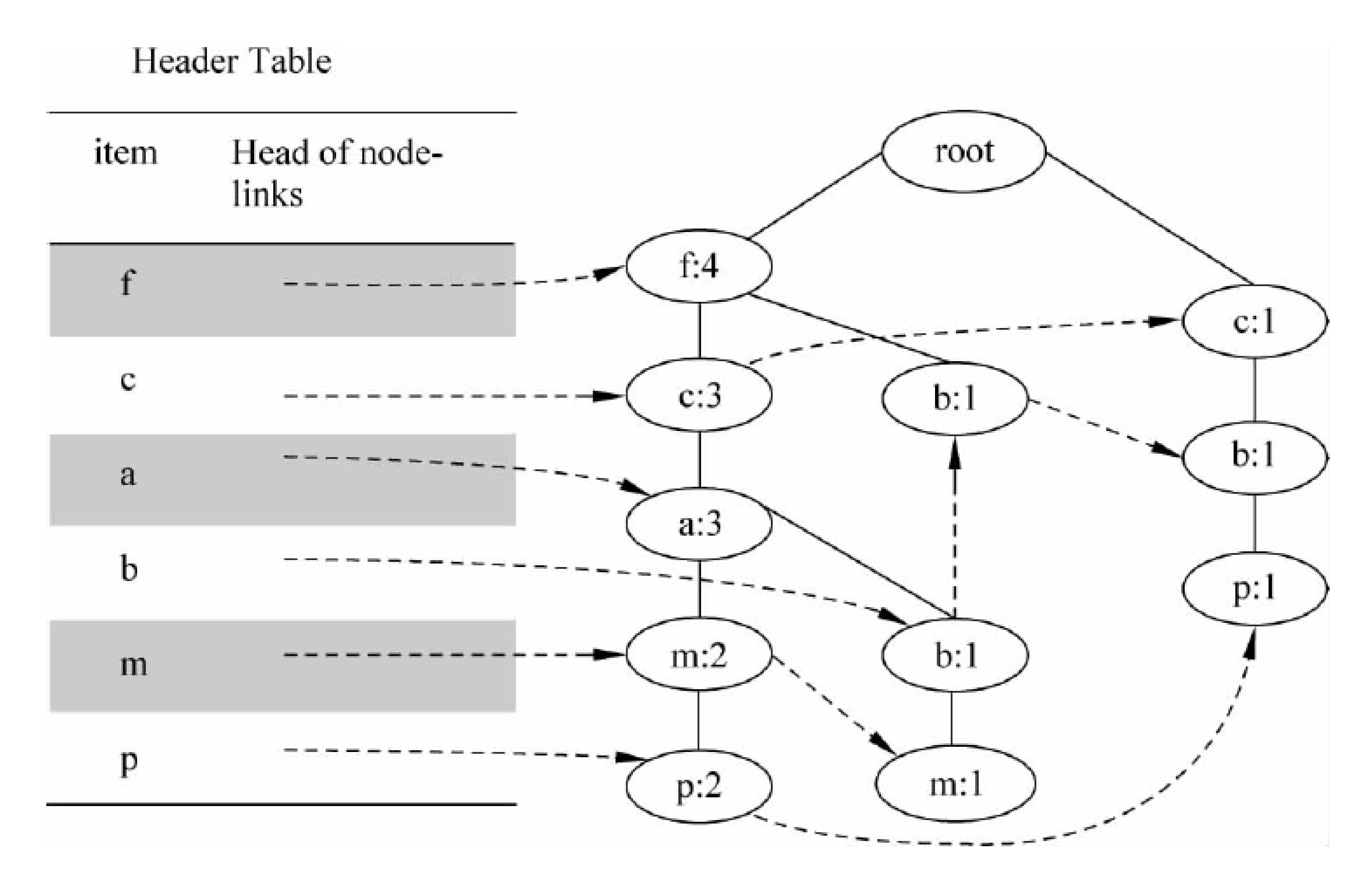


图 7-16 FP-tree 结构图

FP-tree 具有如下性质。

(1) 它有一个根节点,即 root 节点。

(2) 每个节点存储三个信息:项名称,计数,节点链接。

(3) Header Table 中的每一项存储两个信息:项名称和节点链接。其中,节点链接指向该项在 FP-tree 中的第一个节点。

另外,既然在第一次遍历后得到了每个项的频度,那么就可以根据频度对任意一条事务记录(Transaction)排序。下面遍历数据集,构造 FP-tree。

(1) 逐行读取数据集,每行数据由多个项组成;

(2) 对每行数据中的项按频度高低排序;

(3) 把每行数据都插入到 FP-tree 中。

把每行数据插入到 FP-tree 中时,依照以下规则:按频度从高到低依次读取项,将 root 节点设为当前节点,对于每一项,如果项存在于当前节点的子节点中,则将此子节点的计数加 1,并将其设为当前节点,若不存在则创建与项名称相同的子节点并设其计数为 1。

当数据集中的所有数据都被遍历后,FP-tree 建立完成。

接下来就进入挖掘阶段,挖掘的结果就是候选规则。

输入挖掘算法 FP-growth 函数的有两个参数,一个是 FP-tree,另一个是项集 α。第一次调用时 α 为空集。算法执行如下。

若 FP-tree 只有一条路径,则输出这条路径上项的所有组合与 α 的并集,支持度是这个组合中项的计数的最小值。

若 FP-tree 里有不止一条路径,则对于 Header Table 里的每一项 α 进行如下操作。

(1) 输出 α 与 a 的并集 β,其支持度是 α 的支持度;

(2) 建立条件 FP-tree,若条件 FP-tree 非空,则以条件 FP-tree 和 β 为参数调用 FP-growth 函数。

条件 FP-tree 指的是,以在 α 的条件上筛选出来的数据集建立的 FP-tree。

当计算出候选规则及其支持度以后,就可以在相应的置信度基础上挖掘出关联规则了。

7.5.3 回归和分类分析

回归分析是确定两种或两种以上变数之间相互依赖的定量关系的一种统计分析方法,运用十分广泛。回归分析按照涉及的自变量的多少,可分为一元回归分析和多元回归分析;按照自变量和因变量之间的关系类型,可分为线性回归分析和非线性回归分析。如果在回归分析中,只包括一个自变量和一个因变量,且二者的关系可用一条直线近似表示,这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量,且因变量和自变量之间是线性关系,则称为多元线性回归。

分类数据是统计数据的一种,指反映事物类别的数据,如人按性别分为男、女两类。分类数据是离散数据。分类属性具有有限个(但可能更多)不同值,值之间无序。有很多方法产生分类数据的概念分层。分类分析是指找出数据库中的一组数据对象的共同特点并按照分类模式将其划分为不同的类,其目的是通过分类模型,将数据库中的数据项映射到某个给定的类别中。可以应用到涉及应用分类、趋势预测中,如淘宝商铺将用户在一段时间内的购买情况划分成不同的类,根据情况向用户推荐关联类的商品,从而增加商铺的销售量。

回归(Regression)关注的是输入变量和结果之间的关系。“回归”这个术语最早由费兰西斯.高尔顿在19世纪用来描述生物现象。这种现象是拥有较高高度的祖先的后代往往回归到正常的平均水平。具体地说,回归分析有助于了解一个目标变量如何随着属性变量的变化而变化。

例如一些问题:我想预测客户的生命周期价值,并且了解是什么因素在其中产生影响。是什么使得价值更高或更低?我想预测这个贷款是否会被拖欠?

回归分析的结果可以是连续的或离散的,如果是离散的,还可以预测各个离散值产生的概率。

1. 线性回归

本节将介绍回归分析中的一种——线性回归。之前介绍了关联规则分析适用于处理离散型数据,比如电子商务交易记录等,但不适用处理数值型的连续数据。本节介绍的线性回归正是适合处理数值型的连续数据。

线性回归是统计学的一种常用方法,它的主导思想是利用预定的权值将属性进行线性组合来表示类别,如下面公式所示:

$$\chi = w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k$$

式中的 χ 是目标变量,a 是属性值,w 是权值。

线性回归的输出就是权值,即

(1) 一组系数,表示相应的属性值的相对影响;

(2) 一个以线性表达来预测结果的函数。

对于这个目标函数,我们感兴趣的是预测值和真实值的差异。最好的目标是预测值和真实值的差距最小。那么预测值和真实值的差值的平方之和如下:

$$\sum_{i=1}^{n}\left(\chi^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)}\right)^2$$

括号里的表达式是第 i 个示例的真实类值和它的预测类值之差。我们正是需要通过选择适当的系数来使得这个平方和的值最小化。

这里介绍一个离散型变量的例子,例中的模型公式如下:

income = b_0 + $b_1$age + $b_2$yearOfEducation + $b_3$gender + $b_4$state

这个模型用来预测收入(income),影响模型的变量包括:年龄(age),受教育的时间(yearOfEducation),性别(gender)和国家(state)。其中,性别和国家是离散性数据。性别只有“男”或“女”,国家则可能有几十个之多。因为线性回归中的一个假设是收入符合正态分布,但实际上往往不是这样。因此,更好的一个选择是选择收入的对数。

第一个系数 b_0 表示的是当所有的变量都为0时的收入。正如上文所述,线性回归不仅有预测作用,也有解释作用,即能解释各个变量对目标变量的影响程度。假如我们问一个问题:年龄对收入有影响吗?如果答案是否定的,那么age的系数应该是0。

线性回归是一个出色的、简单的、适用于数值预测的方法,在统计应用领域得到了广泛的应用。当然,也存在一定的缺陷。如果数据呈现非线性关系,线性回归将只能到一条“最适合”的直线,“最适合”指的是最小均方差。线性模型也是学习其他更为复杂模型的基础。总之,线性回归的优点和缺点如表7-11所示。

表 7-11　线性回归比较分析

优　　点	缺　　点
准确地表达输入变量与目标变量之间的关系(利用系数)	不能很好地处理缺失值
对冗余的变量具有抗干扰性	假设每个变量对目标变量的影响都是线性的
对变量的影响有解释作用	不能处理以非连续方式影响目标变量的变量
容易对测试数据进行预测	不能很好地处理离散型变量

2. 逻辑回归

逻辑回归是用来预估一个事件发生的几率的模型。一个典型的例子是：通过对贷款人的信用分数、收入、贷款规模等因素进行建模，从而计算出这个贷款人能偿还贷款的几率。逻辑回归也可以被看成是一个分类器，以概率最高的类别来预测。在逻辑回归中，输入变量可以是连续的，也可以是离散的。

逻辑回归是在处理一些二元分类问题时的首选方法，例如，

(1) 真/假；

(2) 批准/拒绝；

(3) 有回应/无回应；

(4) 购买/不购买；

(5) 中国男足是否会赢得下届世界杯。

逻辑回归如图 7-17 所示。

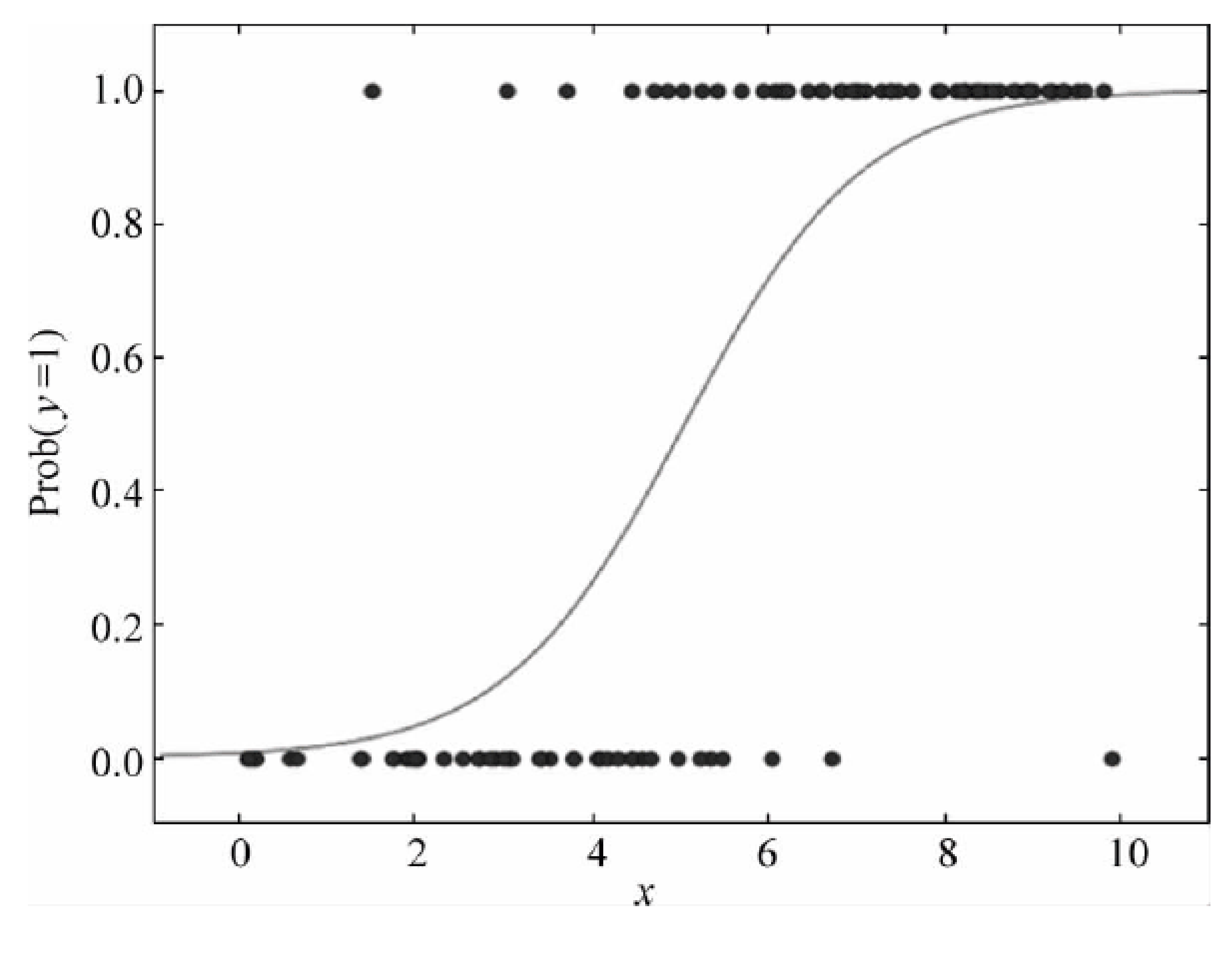


图 7-17　逻辑回归图

所以如果我们不仅对预测类感兴趣，而且对某一类事件发生的概率也感兴趣，那么逻辑回归是特别适合的。下面以贷款的模型来解释逻辑回归，模型公式如下：

```
default = f(creditScore, income, loanAmt, existingDebt)
```

上式表示了通过信用等级(creditScore)、收入(income)、贷款总额(loanAmt)、已有债务(existingDebt)等几个输入来预测贷款人能偿还贷款的概率。这个概率应该是在 0 和 1 之间，1 表示不能偿还，0 表示能偿还。如果需要一个类似“是/否”的答案，那么可以设置一个

阈值0.5。与上文中的线性回归不同,逻辑回归有两个缺陷:第一,目标属性从函数中产生的不是概率值,因为目标关系值有可能落在0~1的范围以外;第二,最小平方回归假设误差不但是统计上的独立,而且呈现出具有相同标准差的正态分布。

从线性回归开始,若想得到一个0~1之间的概率,当目标变量为1时,有

$$\ln \frac{p(y=1)}{1-p(y=1)} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

转换后可以表达为

$$\frac{p(y=1)}{1-p(y=1)} = \exp\left(\sum_{i=1}^{k} b_i x_i\right) = \prod_{i=0}^{k} \exp(b_i x_i)$$

这样,其中就有个很有趣的事实:概率的规模等于数量总和。假如训练数据中13%的数据显示"不能偿还",则所有训练集的得分总和相当于训练例子数量的13%。假如贷款申请者中收入小于5万元的有40%会拖欠,则在这个收入分类中的人训练集得分总和是这个分类中例子数量的40%。

小结:逻辑回归具有可解释性的输出值,而且可以很简单地确定变量影响的结果。这使得它比线性回归也更复杂一些。同样,它对冗余的变量也具有稳定性,对输出也有准确的表示,容易对测试数据进行预测。逻辑回归不仅返回某个事件发生的概率,而且保留了训练数据的一些统计信息。当然,逻辑回归也具有线性回归的缺点。它不能很好地处理缺失值,仍然默认变量是以线性的方式影响结果。所以,如果要把它应用到一些非线性关系的问题中,那么模型就有一定的局限性。总之,逻辑回归的优缺点如表7-12所示。

表7-12 逻辑回归比较分析

优　　点	缺　　点
可解释的结果	不能很好地处理默认值
对冗余变量具有干扰性	假设变量都是以线性的方式影响结果
对系数有准确的表示	不能处理以非连续方式影响结果的变量
容易对测试数据进行预测	不能很好地处理离散性变量
返回的是一个事件的概率	
保留了统计数据中的统计信息	

3. 多项式回归

对于一个回归方程,如果自变量的指数大于1,那么它就是多项式回归方程。在这种回归技术中,最佳拟合线不是直线,而是一个用于拟合数据点的曲线,如图7-18所示。

重点:虽然会有一个诱导可以拟合一个高次多项式并得到较低的错误,但这可能会导致过拟合。需要经常画出关系图来查看拟合情况,并且专注于保证拟合合理,既没有过拟合又没有欠拟合。图7-19是一个图例,可以帮助理解。

明显地向两端寻找曲线点,看看这些形状和趋势是否有意义。更高次的多项式最后可能产生怪异的推断结果。

4. 逐步回归

在处理多个自变量时,可以使用这种形式的回归。在这种技术中,自变量的选择是在一个自动的过程中完成的,其中包括非人为操作。这一壮举是通过观察统计的值,如R-

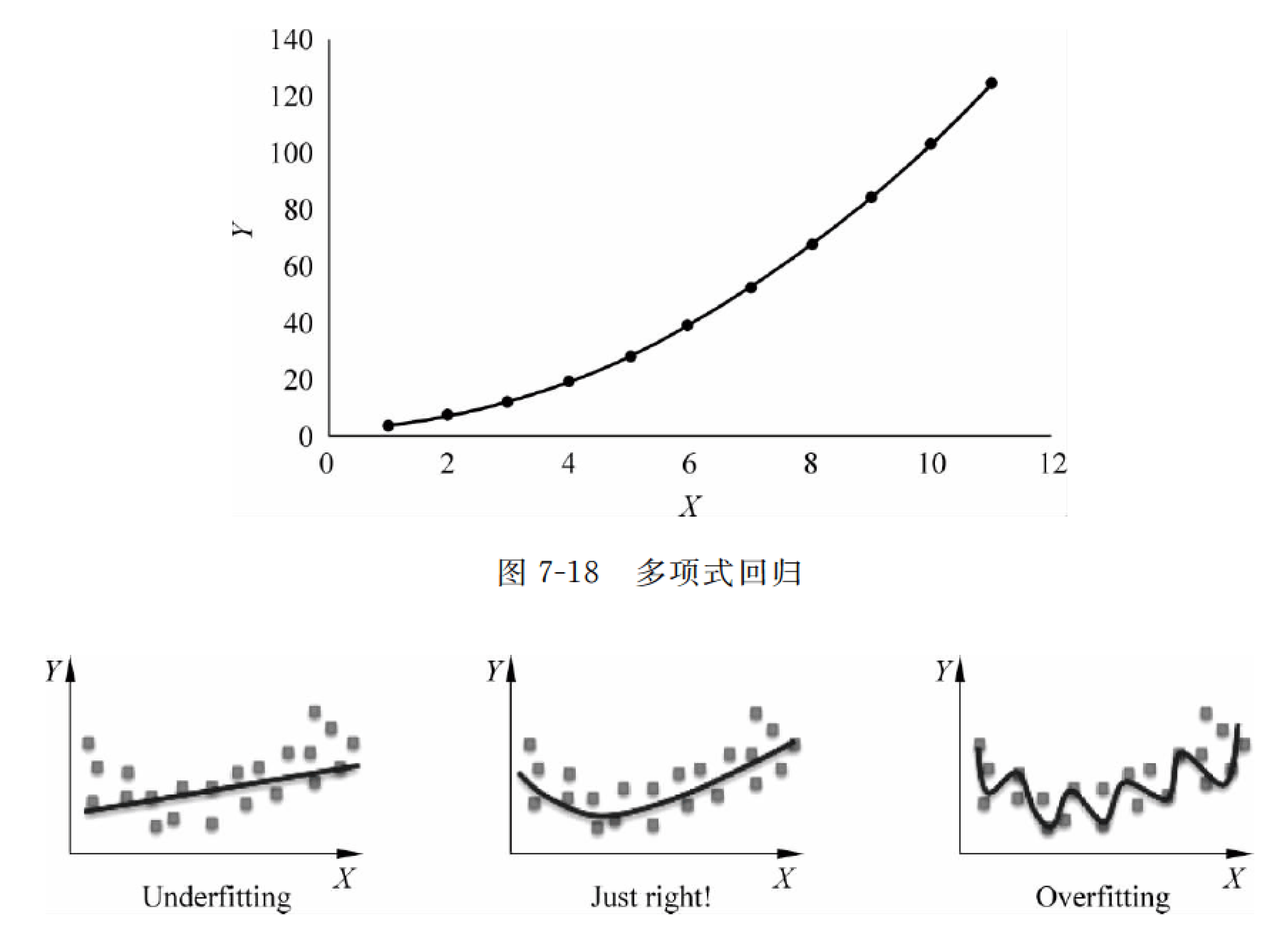


图 7-18　多项式回归

图 7-19　多项式回归拟合比较

square、t-status 和 AIC 指标，来识别重要的变量。逐步回归通过同时添加/删除基于指定标准的协变量来拟合模型。下面列出了一些最常用的逐步回归方法：标准逐步回归法做两件事情，即增加和删除每个步骤所需的预测；向前选择法从模型中最显著的预测开始，然后为每一步添加变量；向后剔除法与模型的所有预测同时开始，然后在每一步消除最小显著性的变量。

5. 朴素贝叶斯

分类问题中的主要任务是预测目标所属的类别。与聚类不同的是，这里类别的种类是事先已经定义好的。

朴素贝叶斯分类器(Naïve Bayesian Classifier)是一个简单的基于贝叶斯理论的概率分类器。朴素贝叶斯分类器假设属性之间相互独立。或者说，一个朴素贝叶斯分类器假设某个类的特性的出现与其他特征没有关系。虽然这个假设在实际应用中往往是不成立的，但朴素贝叶斯分类器依然有着坚实的数学基础、稳定的分类效率。

比如，一个物体可以依据它的形状、大小、颜色等属性被分类成某个类别(网球是圆的、直径 6cm、黄颜色)。即使这些属性之间互相有依赖关系存在，朴素贝叶斯分类器也会认为所有的属性之间是无关的。

根据概率模型的特征，朴素贝叶斯分类器可以在有监督的环境下有效地被训练。贝叶斯理论被广泛地应用到文本分类中，例如，可以回答如下问题。

(1) 这封邮件是垃圾邮件吗?

(2) 这名政客属于民主党派还是共和党派?

(3) 网页内容的主题分类有哪些?

通常的朴素贝叶斯模型中，输入变量都是离散型的，当然也有一些算法的变种用来处理

连续型变量。算法的输出是概率的打分，通常是 0～1 之间，可以根据概率最高的类来做预测。

贝叶斯定理是朴素贝叶斯模型的基础，是由英国数学家贝叶斯(Thomas Bayes)的名字命名的。贝叶斯定理是用来描述两个条件概率之间的关系，比如，$P(A|B)$和 $P(B|A)$。贝叶斯规则的描述公式为

$$P(B \mid A)=\frac{P(A \cap B)}{P(A)}=\frac{P(A \mid B)P(B)}{P(A)}$$

贝叶斯公司说明：A 为真的条件下 B 也为真的概率乘以 A 为真的概率，等于 B 为真时 A 为真的概率乘以 B 为真的概率。如果 $P(B|A)$是观察到某个特定类标签时的后验概率，给定我们观察到的一个变量 A，那么依据贝叶斯公式，这个概率与类别 B 中观察到 A 的概率乘以在类别 B 中的先验概率相等。

贝叶斯之所以重要的原因是我们不知道 $P(B|A)$，并且这就是我们要知道的。在我们从训练数据中知道 $P(A|B)$和 $P(B)$的情况下，通常我们并不知道 $P(A)$。下面通过一个例子来说明。

A 同学经常要乘坐飞机在各地间出差，并且将自己的机票升级为头等舱。

A 同学发现，假如他在起飞前至少两小时办理登机手续，那么他能成功把机票升级到头等舱的几率是 75%；反之，他升级到头等舱失败的概率是 35%。在 A 同学忙碌的行程安排中，他只有 40%的时间能赶在起飞前两小时办理登机手续。

假设 A 同学在最近的一次尝试中没能够升级到头等舱的机票，那么他到机场晚于起飞前两小时的概率是多少？这里设定：

X——表示他到机场晚了；

Y——表示他没能订到头等舱机票。

$P(X)$＝到机场晚了的先验概率＝1－0.4＝60%

$P(Y)$＝没能订到头等舱的先验概率＝1－(0.4×0.75＋0.6×0.35)＝1－0.51＝49%

$P(Y|X)$＝在到机场晚了的情况下没有能升级到头登舱的概率＝1－0.35＝65%

可以得到，$P(X|Y)$＝没能升级到头等舱的情况下到机场晚了的概率＝$[P(Y|X)\times P(X)]/P(Y)$＝(0.65×0.6)/0.49≈80%

通过贝叶斯公式，可以得到贝叶斯分类器。假如对于属性 A 有 m 个类别，$a_1,a_2,\cdots,a_m$，那么在给定 j 个 b 变量值的情况下，A 的概率是各个给定 b_j 时的 a_i 的条件概率的乘积，如下式所示：

$$P(A \mid b_j)=P(a_1,a_2,\cdots,a_m \mid b_j)=\prod_{i=1}^{m} P(a_i \mid b_j)$$

于是，有

$$P(b_j \mid a_1,a_2,\cdots,a_m)=\frac{\prod_{i=1}^{m} P(a_i \mid b_j)P(b_j)}{P(a_1,a_2,\cdots,a_m)}$$

因为贝叶斯公式中的条件独立假设，所以上式的分母为 1，可以去掉。

那么，要训练一个朴素贝叶斯分类器，只需要搜集以下的统计数据。

(1) 所有类标签的概率。例如，所有好信用(credit_good)的概率和坏信用(credit_bad)

的概率，从已有的训练数据集中可以得到 $P(\text{good})=0.7$ 和 $P(\text{bad})=0.3$。

(2) 训练数据中有多个属性，对于每个变量和类标签的组合，要计算出它们的条件概率。例如，拥有房子与好信用(own_house/credit_good)，拥有房子与坏信用(own_house/credit_bad)，熟练工作与好信用(job_skilled/credit)，熟练工作与坏信用(job_skilled/credit_bad)。

在计算完各个类的概率及各个属性在给定类下的条件概率后，就可以计算这两者的乘积了。若要对新的数据分配类标签，只需要计算出它在哪个类的打分最高即可，公式如下：

$$\prod_{i=1}^{m} P(a_i \mid b_j)P(b_j)$$

再以上文的信用作为例子，假如计算出各种情况的概率如表 7-13 所示。

表 7-13　贝叶斯模型运算表

a_i	b_j	$P(a_i \mid b_j)$
female	good	0.28
female	bad	0.36
own	good	0.75
own	bad	0.62
Self emp	good	0.14
Self emp	bad	0.17
Savings > 1000	good	0.06
Savings > 1000	bad	0.02

表格中的属性分别表示性别、拥有房产、个体经营以及存款大于 1000 元。

然后，我们有个需要做出判断的例子。X 表示一位女士、拥有房产、个体经营，且存款大于 1000。如何对她做出判断呢？她的信用是好(good)还是坏(bad)呢？

在建立了分类器后，可以找到 $P(\text{good} \mid X)=0.0012$，而 $P(\text{bad} \mid X)=0.0002$。这两者中的最大值被用来分类，即可以判断这名女士的信用记录是好(good)。

到此，我们介绍了朴素贝叶斯分类器以及应用朴素贝叶斯分类的例子，可以看到，朴素贝叶斯分类器的优点和缺点如表 7-14 所示。

表 7-14　贝叶斯分类比较分析

优　　点	缺　　点
能够很好地处理缺失值	数值型变量会被转成离散型
对不相关的变量具有抗干扰性	对相关变量很敏感(不符合条件独立假设)
实现简单	不适用于估计概率
对数据的打分简单	
对过拟合有抵抗性	
处理高纬度的问题时计算效率高	

6. 决策树

决策树是一种非常常见且灵活的用来开发数据挖掘应用的方法。

(1) 分类树用于将要预测的数据划分到同质的组中(分配类标签)。通常应用于二分或

多类别的分类。

(2) 回归树是回归的变种,通常每个节点返回的是目标变量的平均值。回归树通常被应用于连续性数据的分类,比如账户支出或个人收入。

决策树的输入值可以是连续的也可以是离散的,输出是一个用来描述决策流程的树状模型。决策树的叶子节点返回的是类标签或者是类标签的概率分数。理论上,决策树可以被转换成类似上文关联规则中的规则。

因为决策树可以应用到各种不同的情境中,所以决策树应用比较广。决策树的分类规则也很直接,结果也很容易被可视化展现。另外,因为决策树的决策结果是一系列的"如果……就……"表达式,所以决策树的模型中没有隐含的假设,比如,依赖变量和目标变量之间的线性或非线性关系。

决策树通常以流程图的形式展现,如图7-20所示。

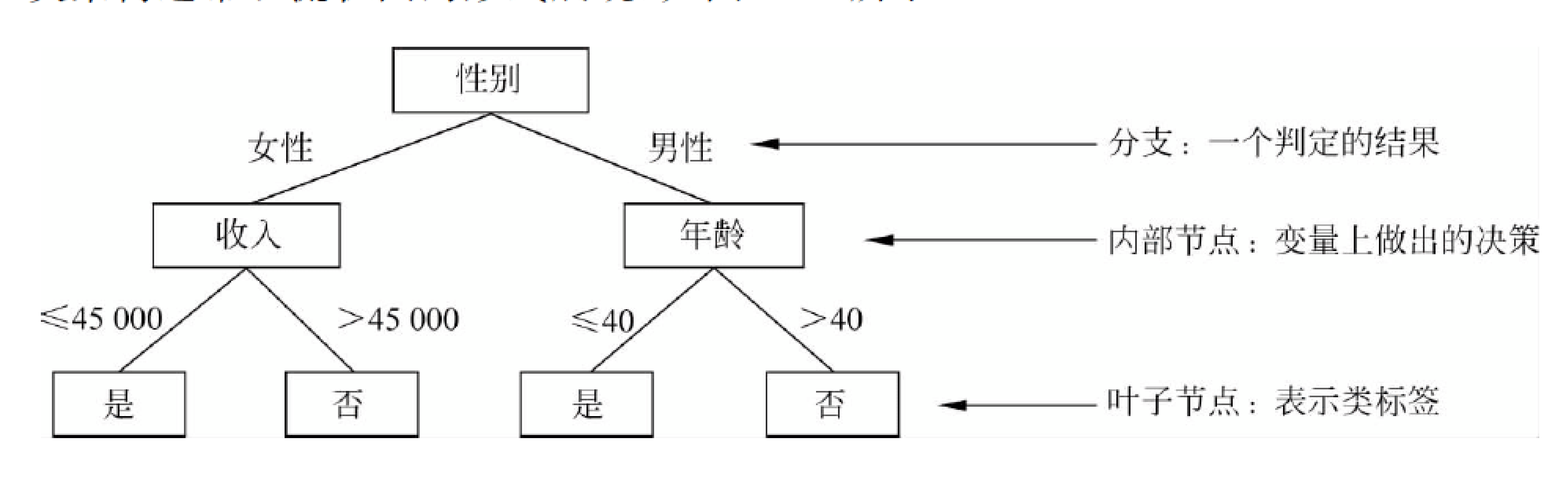


图7-20 决策树流程图

分支:指的是一个决策做出的结果,以连续的方式展现。如果是数值型变量,可以根据变量的不同,将"等于"放在左分支或右分支。

内部节点:指的是决策树内部用来做决策的节点。每个节点对应一个变量或属性。尽管图中展示的是有两个决策结果的节点,但一个节点可以有超过两个的分支。

叶子节点:指的是分支的终点,表示所有之前的决定产生的一个结果。

那么,决策树是如何被构建起来的?

找出"最具有信息"的属性有很多方式,这里介绍一种基于熵的方式,其公式如下:

$$H = -\sum_{c} p(c)\log_2 p(c)$$

其中,$p(c)$是某个类标签c的概率。从上式可以看出,当$p(c)=0$或者1时,$H=0$。所以对于一个二元分类问题,$H=0$意味着节点很"纯净"。当每个类的可能性都相等的时候,H的值最大。

接着,我们可以找到条件熵。条件熵指的是每个属性的类标签的熵的权重和,其计算公式如下:

$$H = -\sum_{v} p(v) \sum_{c} p(c \mid v)\log_2 p(c \mid v)$$

假如有个属性"住房",这个属性有三种值"无房""租房""有房"。直观上,有房的人的信用度应该比租房的高一些,租房的信用度应该比无房的高一些。所以"住房"这个属性可以对类标签的划分给出更多的信息,属于"更有信息"的属性。比如,已知一个住房信息如表7-15所示。

表 7-15　住房信息

住房属性	无房	有房	租房
$P(\text{housing})$	0.108	0.713	0.179
$P(\text{bad}\mid\text{housing})$	0.407	0.261	0.391
$P(\text{good}\mid\text{housing})$	0.592	0.739	0.601

那么有

$$\begin{aligned} H_{(\text{housing}\mid\text{credit})} &= -[0.108\times(0.407\log_2(0.407)+0.592\log_2(0.592)) \\ &\quad +0.713\times(0.261\log_2(0.261)+0.739\log_2(0.739)) \\ &\quad +0.179\times(0.391\log_2(0.391)+0.601\log_2(0.601))] \\ &= 0.868 \end{aligned}$$

就可以计算出类标签在“住房”属性上的条件熵。

计算出熵后，就可以挑选出“最有信息”的属性了，按照下式计算出信息增益(InfoGain)值：

$$\text{InfoGain}_{\text{housing}} = H_{\text{credit}} - H_{\text{housing}\mid\text{credit}} = 0.88 - 0.86 = 0.02$$

所有属性中，信息增益值最高的属性就是要找的属性。依次就可递归地构建决策树了。

总结：决策树既能够处理数值型数据，也能够处理类别型数据，是一种很强大的数据挖掘工具。当变量之间的关系不是线性关系时，线性回归模型就不能正确处理数据了，但是决策树不存在这样的问题。决策树具有高效率地计算以及打分简单的特点，输出结果容易理解。但是决策树对训练数据中的很小的变化很敏感，假如有一个很大的数据集，用其中的两个不同子集建立两个决策树，会发现它们的差距很大，即使它们来自同一个数据集。如果决策树建得过深，又容易导致过拟合问题。决策树的优缺点如表 7-16 所示。

表 7-16　决策树比较分析

优　　点	缺　　点
输入类型不受限制	树结构对训练集的细小改变很敏感
对冗余的、相关联的变量具有抗干扰性	树构造得过深容易导致过拟合
自然地处理变量之间的关系	不适应用于依赖多个变量的结果
能处理具有非线性关系的变量	不能很好地处理缺失值
构建的效率高	实际中，决策的规则可能比较复杂
容易对测试数据进行分类	
很多算法能返回变量权限值上的度量	

7. 随机森林

在分布式环境中，通常节点要独立地进行计算，且分布式环境中最稀缺的资源是网络。这样的情况下，训练一个决策树是比较困难的，一种更好的办法是利用集成学习的方法。对于决策树，可以在分布式环境中独立地训练多个决策树，利用多个决策树来分类，最后把结果聚集起来。利用多个决策树来分类的方法叫“随机森林”。

随机森林是由 Breiman 于 2001 年提出来的，它是一个包含多个决策树的分类器，其输出的类别由树输出的类别的众数而定。为了构建多个不同的决策树，随机森林采用从数据

中随机抽样的方法。在之前决策树的构建中，每个节点的产生是通过挑选变量来生成节点，但是在随机森林中，是从原始数据中随机抽样出来的子集来训练决策树。直观上，这样的做法可能有点儿违反直觉，但是，“随机森林”在与许多其他分类器相比较时取得了很好的效果，包括支持向量机（SVM）、神经网络（Neural Network）等，而且对过拟合问题有很好的规避。另外，“随机森林”也是对用户很友好的工具，因为它只有两个参数：随机抽样子集上的变量的数目、树的数量。而且随机森林对这两个参数也不敏感。下面介绍“随机森林”算法。

（1）从原始数据中产生 n 个随机抽样。

（2）对于每一个抽样，训练一个未剪枝的决策树或回归树：对于每个节点，不是在所有的属性中挑选分割得最好的属性，而是在 m 个抽样出来的属性中挑选出分割得最好的那个。

（3）对数据进行预测，并搜集各个树的预测结果，以众数（出现次数最多的值）给出最后的预测结果。

当“随机森林”构建好以后，要对这个模型的错误率进行衡量，以了解这个模型的准确度。可以使用一种 OOB（Out-Of-Bag）的估计来衡量。

（1）对于每一次迭代抽样，用抽样的数据来训练模型，用不在抽样中的数据来预测。

（2）搜集所有 OOB 预测的平均值，计算错误率。

通常，当有足够多的树被生成的时候，OOB 的估计结果会比较准确。

使用“随机森林”的方法时，需要了解以下几点。

（1）若想获得好的效果，树的数目就有必要随着属性的增加而增加。决定应该训练多少棵树的最好的方式是将一个森林的预测结果和它的子集的预测结果做比较。当子集的结果和整个森林的预测结果一样好时，那么树的数量就足够了。

（2）当挑选 m 个变量时，Breiman 教授建议分别尝试将参数折半或翻倍，然后从中挑选出最好的。

（3）若想得到一个稳定的变量权值与距离的估计，就有必要训练很多棵树。

（4）当训练数据中各个类标签的比例不平衡甚至差别很大时，就有必要改变预测时的规则（不一定由众数决定）。

（5）当训练数据集比较大或者要训练的树的数量比较多的时候，可以在同一时刻只在内存保留一棵树，这样可以更节约内存。

“随机森林”优缺点比较如表 7-17 所示。

表 7-17 “随机森林”比较分析

优　　点	缺　　点
准确率很高	在一些有噪声的分类/回归任务上，“随机森林”被观察到有过拟合问题
在大数据集上的计算效率很高	对于有不同级别属性的数据，级别划分较多的属性会对“随机森林”产生更大的影响
能够很好地处理数量很多的输入变量，比如，上千个输入变量	
它能给出变量在分类中的权值的一个估计	
在森林构造的过程中，它能够产生一个内部的无偏见的泛化误差的估计	

续表

优　　点	缺　　点
它对于处理没有初始值的数据是有效的方法，即使当数据有大部分缺失的时候还能保持准确度	
对于类标签不平衡的训练数据，它也是一种很好的手段	
它在分布式环境上有很好的伸缩性	

7.5.4 时序模型

在生产和科学研究中，对某一个或一组变量 $x(t)$ 进行观察测量，将在一系列时刻 t_1，t_2，…，t_n（t 为自变量且 $t_1<t_2<\cdots<t_n$）所得到的离散数字组成序列集合 $x(t_1)$，$x(t_2)$，…，$x(t_n)$，我们称之为时间序列。这种有时间意义的序列也称为动态数据。这样的动态数据在自然、经济及社会等领域都是很常见的。如在一定生态条件下，动植物种群数量逐月或逐年的消长过程、某证券交易所每天的收盘指数、每个月的GNP、失业人数或物价指数等。

时间序列分析是根据系统观测得到的时间序列数据，通过曲线拟合和参数估计来建立数学模型的理论和方法。它一般采用曲线拟合和参数估计方法（如非线性最小二乘法）进行，如图7-21所示。时间序列分析常用在国民经济宏观控制、区域综合发展规划、企业经营管理、市场潜量预测、气象预报、水文预报、地震前兆预报、环境污染控制、生态平衡、天文学和海洋学等方面。

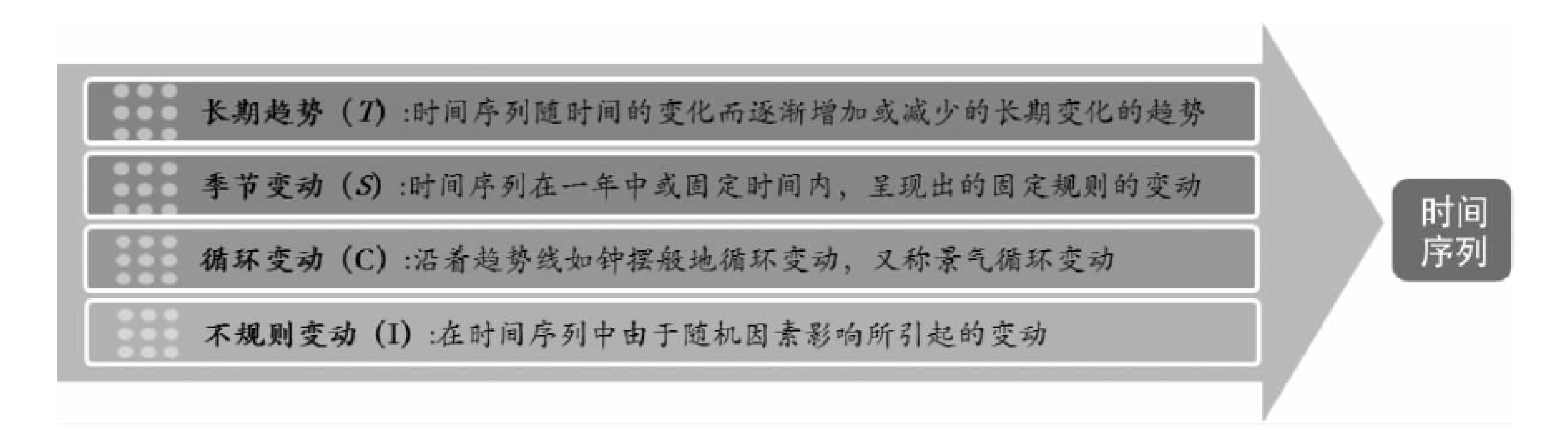


图 7-21　时间序列构成

时间序列表示通常分为两个步骤：①先形成散点图；②根据图形走势选择合适的模型。时间序列模型主要有直线型、指数型、二次抛物线型或组合模型等。组合模型分为加法模型和乘法模型。

（1）加法模型：假定时间序列是基于4种成分相加而成的。长期趋势并不影响季节变动；$Y=T+S+C+I$。

（2）乘法模型：假定时间序列是基于4种成分相乘而成的。假定季节变动与循环变动为长期趋势的函数。

ARMA模型的全称是自回归移动平均模型（Auto Regression Moving Average Model），它是目前最常用的拟合平稳序列的模型，又可细分为AR模型（Auto Regression Model）、MA模型（Moving Average Model）和ARMA模型（Auto Regression Moving Average Model）。ARIMA模型（Auto Regressive Integrated Moving Average Model）又称自回归求和移动平均模型，当时间序列本身不平稳的时候，如果它的增量，即的一次差分，稳定在零点附近，可以将它看成是平稳序列。在实际的问题中，所遇到的多数非平稳序列可以通

过一次或多次差分后成为平稳时间序列，则可以建立模型。这说明任何非平稳序列只要通过适当阶数的差分运算实现差分后平稳，就可以对差分后序列进行 ARIMA 模型拟合了。ARIMA(p,d,q)模型是指 d 阶差分后自相关最高阶数为 p，移动平均最高阶数为 q 的模型，通常它包含 $p+q$ 个独立的未知系数。

(1) AR(p)(p 阶自回归模型)。

$$x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + u_t$$

其中，u_t 是白噪声序列，δ 是常数(表示序列数据没有 0 均值化)。

$$\text{AR}(p)\ 等价于 (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) x_t = \delta + u_t$$

AR(p)的特征方程是

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p = 0$$

AR(p)平稳的充要条件是特征根都在单位圆之外。

(2) MA(q)(q 阶移动平均模型)。

$$x_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$
$$x_t - \mu = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q) u_t = \Theta(L) u_t$$

其中，$\{u_t\}$是白噪声过程。

MA(q)平稳性：

MA(q)是由 u_t 本身和 q 个 u_t 的滞后项加权平均构造出来的，因此它是平稳的。

MA(q)可逆性(用自回归序列表示 u_t)

$$u_t = [\Theta(L)]^{-1} x_t$$

可逆条件：即$[\Theta(L)]^{-1}$收敛的条件。即 $\Theta(L)$每个特征根绝对值大于 1，即全部特征根在单位圆之外。

(3) ARMA(p,q)(自回归移动平均过程)。

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \delta + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$
$$\Phi(L) x_t = (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) x_t$$
$$= \delta + (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q) u_t = \delta + \Theta(L) u_t$$
$$\Phi(L) x_t = \delta + \Theta(L) u_t$$

ARMA(p,q)平稳性的条件是方程 $\Phi(L)=0$ 的根都在单位圆外；可逆性条件是方程 $\Theta(L)=0$ 的根全部在单位圆外。

(4) ARIMA(p,d,q)(单整自回归移动平均模型)。

差分算子

$$\Delta x_t = x_t - x_{t-1} = x_t - L x_t = (1 - L) x_t$$
$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1} = (1 - L) x_t - (1 - L) x_{t-1} = (1 - L)^2 x_t$$
$$\Delta^d x_t = (1 - L)^d x_t$$

对 d 阶单整序列 $x_t \sim I(d)$

$$w_t = \Delta^d x_t = (1 - L)^d x_t$$

则 w_t 是平稳序列，于是可对 w_t 建立 ARMA(p,q)模型，所得到的模型称为 $x_t \sim$ ARIMA(p,d,q)，模型形式是

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + \delta + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$
$$\Phi(L) \Delta^d x_t = \delta + \Theta(L) u_t$$

由此可转化为ARMA模型。

7.5.5 结构优化

用于分析多个变量间可能会存在较多的信息重复，若直接用来分析，会导致模型复杂，同时可能会引起模型较大误差，因此要初步分析数据间的相关性，剔除重复因素。

遗传算法是计算机科学人工智能领域中用于解决最优化的一种搜索启发式算法，是进化算法的一种。这种启发式通常用来生成有用的解决方案来优化和搜索问题。进化算法最初是借鉴了进化生物学中的一些现象而发展起来的，这些现象包括遗传、突变、自然选择以及杂交等，如图7-22所示。

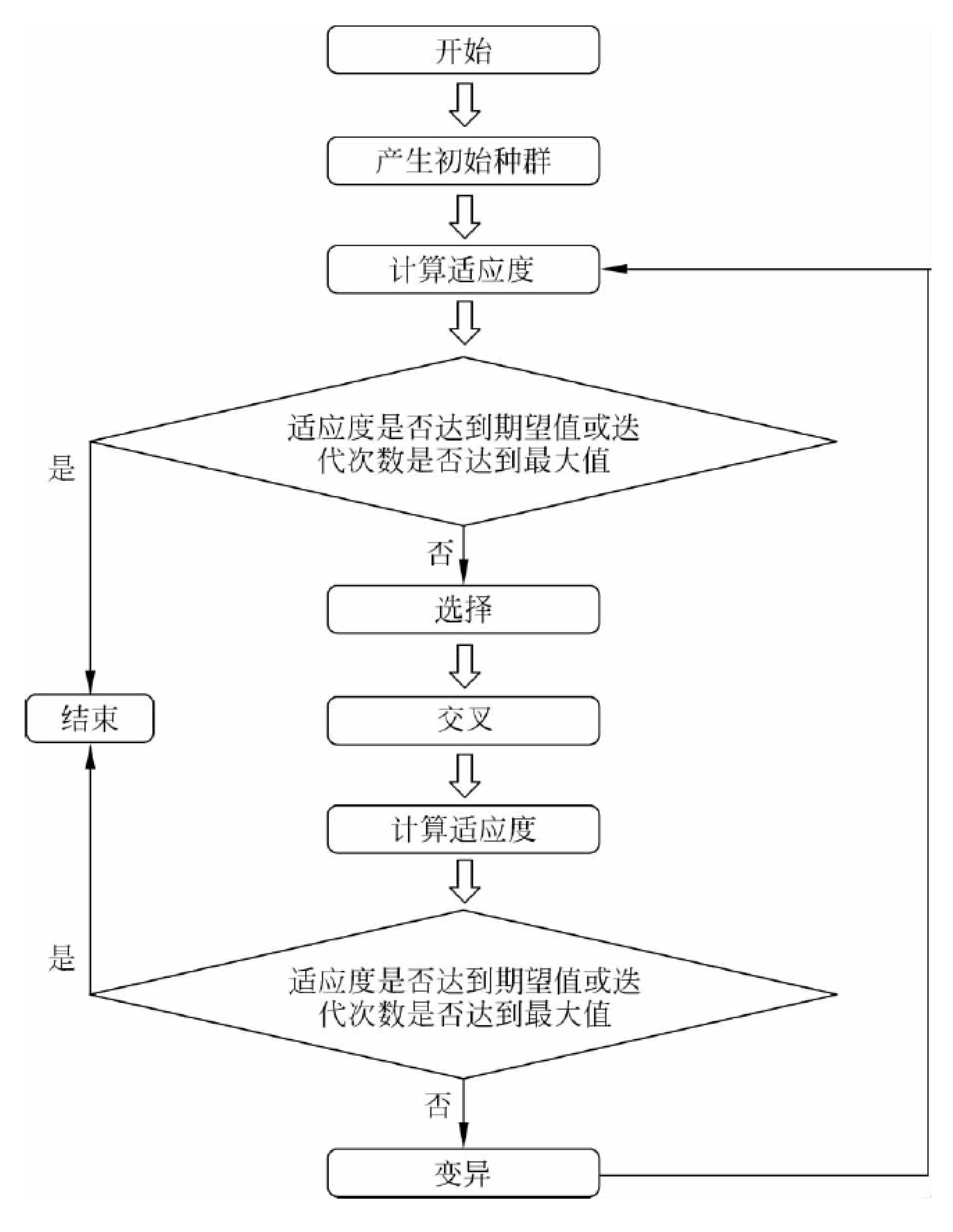


图7-22 遗传算法过程图

遗传算法广泛应用在生物信息学、系统发生学、计算科学、工程学、经济学、化学、制造、数学、物理、药物测量学和其他领域之中。

1. 算法特点

（1）遗传算法从问题解的串集开始搜索，而不是从单个解开始。这是遗传算法与传统优化算法的极大区别。传统优化算法是从单个初始值迭代求最优解的；容易误入局部最优解。遗传算法从串集开始搜索，覆盖面大，利于全局择优。

（2）遗传算法同时处理群体中的多个个体，即对搜索空间中的多个解进行评估，减少了

陷入局部最优解的风险，同时算法本身易于实现并行化。

（3）遗传算法不是采用确定性规则，而是采用概率的变迁规则来指导它的搜索方向。

（4）具有自组织、自适应和自学习性。遗传算法利用进化过程获得的信息自行组织搜索时，适应度大的个体具有较高的生存概率，并获得更适应环境的基因结构。

灰色系统是指“部分信息已知，部分信息未知”的“小样本”，“贫信息”的不确定性系统。它通过对“部分”已知信息的生成、开发去了解、认识现实世界，实现对系统运行行为和演化规律的正确把握和描述。

严格来说，灰色系统是绝对的，而白色与黑色系统是相对的。社会、经济、农业等系统的预测都属于特征性灰色系统的预测。

灰色系统认为：尽管客观系统表象复杂，数据离散，但它们总是有整体功能的，总是有序的。因此，它必然潜藏着某种内在规律。关键在于要用适当方式去挖掘它，然后利用它。

2. 应用

（1）数列预测：即用观察到的反映预测对象特征的时间序列来构造灰色预测模型，预测未来某一时刻的特征量，或达到某一特征量的时间。

（2）灾变与异常值预测：即通过灰色模型预测异常值出现的时刻，预测异常值什么时候出现在特定时区内。

（3）季节灾变与异常值预测：通过灰色模型预测灾变值发生在一年内某个特定的时区或季节的灾变预测。

（4）拓扑预测：将原始数据作曲线，在曲线上按定值寻找该定值发生的所有时点，并以该定点为框架构成时点序列，然后建立模型预测该定值所发生的时点

（5）系统预测：通过对系统行为特征指标建立一组相关联的灰色模型，预测系统中众多变量间的相互协调关系的变化，如图 7-23 所示。

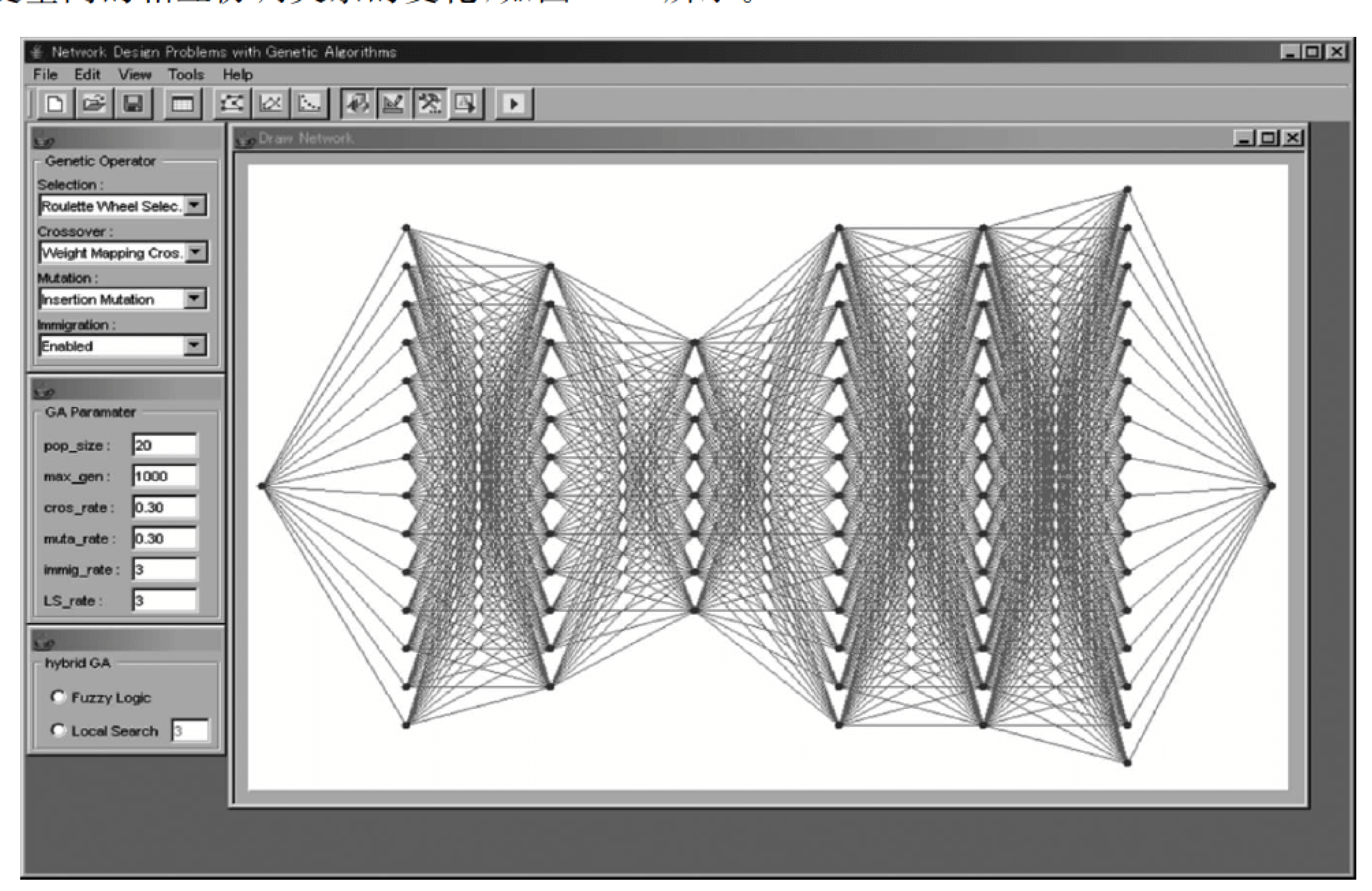


图 7-23　灰色系统模型可视化

7.5.6 深度机器学习

1. 人工神经网络

人工神经网络(Artificial Neural Network,ANN),也被称为神经网络(Neural Network,NN),是一种模拟生物神经网络的数学模型和计算模型。一个神经网络由一组相互关联的人工神经元组成,并通过模拟真实的生物神经系统的交互作用完成计算的过程。在大多数情况下,神经网络是一个自适应系统,通过在学习阶段改变外部输入信息达到改变神经网络结构的目的。按照对生物神经系统模拟时不同的组织层次和抽象层次,神经网络可以分为:神经元层次神经网络、组合式神经网络、网络层次神经网络、神经系统神经网络和智能型神经网络。按照神经网络的结构和学习方式,神经网络可以分为:前馈型神经网络、反馈型神经网络、连续型神经网络和离散型神经网络等。

神经网络的工作方式主要包括两个时期:学习期和工作期。在学习期,神经网络的计算单元状态不变,网络连接权值使用样本学习等方法进行修改;在工作期,神经网络连接权值固定,而各计算单元发生变化,对不同的输入达到稳定状态。目前人工神经网络已经广泛地应用到:函数逼近、概率估计、知识提取、模型分类、数据聚类和最优化计算等领域。

2. 支持向量机

支持向量机(Support Vector Machine,SVM)是一种有监督的学习方法,该方法能够通过分类和回归模型来分析数据和识别模式。SVM 方法是 20 世纪 90 年代初 Vapnik 等人根据统计学习理论提出的一种新的机器学习方法,它以结构风险最小化原则为理论基础,通过适当地选择函数子集及该子集中的判别函数,使学习机器的实际风险达到最小,保证了通过有限训练样本得到的小误差分类器,对独立测试集的测试误差仍然较小。

支持向量机的基本思想是:首先,在线性可分情况下,在原空间寻找两类样本的最优分类超平面。在线性不可分的情况下,加入了松弛变量进行分析,通过使用非线性映射将低维输入空间的样本映射到高维属性空间使其变为线性情况,从而使得在高维属性空间采用线性算法对样本的非线性进行分析成为可能,并在该特征空间中寻找最优分类超平面。其次,它通过使用结构风险最小化原理在属性空间构建最优分类超平面,使得分类器得到全局最优,并在整个样本空间的期望风险以某个概率满足一定上界。

其突出的优点表现在:①基于统计学习理论中结构风险最小化原则和 VC 维理论,具有良好的泛化能力,即由有限的训练样本得到的小的误差能够保证使独立的测试集仍保持小的误差。②支持向量机的求解问题对应的是一个凸优化问题,因此局部最优解一定是全局最优解。③核函数的成功应用,将非线性问题转化为线性问题求解。④分类间隔的最大化,使得支持向量机算法具有较好的鲁棒性。由于 SVM 自身的突出优势,因此被越来越多的研究人员作为强有力的学习工具,以解决模式识别、回归估计等领域的难题。

3. 马尔可夫聚类算法

马尔可夫聚类算法(the Markov Cluster Algorithm,MCL)是图聚类方法的一种,该算法核心的步骤是:使用一个随机过程访问密集群集,直到随机访问所有的顶点之后退出访问这个群集。然而马尔可夫聚类算法不是实际的模拟随机访问过程,而是人为不断地修改访问矩阵的转移概率值。马尔可夫聚类算法的伪代码如下。

(1) sparse autoencoder。deep learning 领域比较重要的一类算法——sparse autoencoder，即稀疏模式的自动编码。sparse autoencoder 是一种自动提取样本(如图像)特征的方法。把输入层激活度(如图像)用隐层激活度表征，再把隐层信息在输出层还原。这样隐层上的信息就是输入层的一个压缩过的表征，且其信息熵会减小。并且这些表征很适合作分类器。我们知道，deep learning 也叫做无监督学习，所以这里的 sparse autoencoder 也应是无监督的。如果是有监督的学习的话，在神经网络中，只需要确定神经网络的结构就可以求出损失函数的表达式了(当然，该表达式需对网络的参数进行“惩罚”，以便使每个参数不要太大)，同时也能够求出损失函数偏导函数的表达式，然后利用优化算法求出网络最优的参数。应该清楚的是，损失函数的表达式中，需要用到有标注值的样本。那么这里的 sparse autoencoder 为什么能够无监督学习呢？难道它的损失函数的表达式中不需要标注的样本值(即通常所说的 y 值)么？其实在稀疏编码中“标注值”也是需要的，只不过它的输出理论值是本身输入的特征值 x，其实这里的标注值 $y=x$。这样做的好处是，网络的隐含层能够很好地代替输入的特征，因为它能够比较准确地还原出那些输入特征值。sparse autoencoder 的一个网络结构如图 7-24 所示。

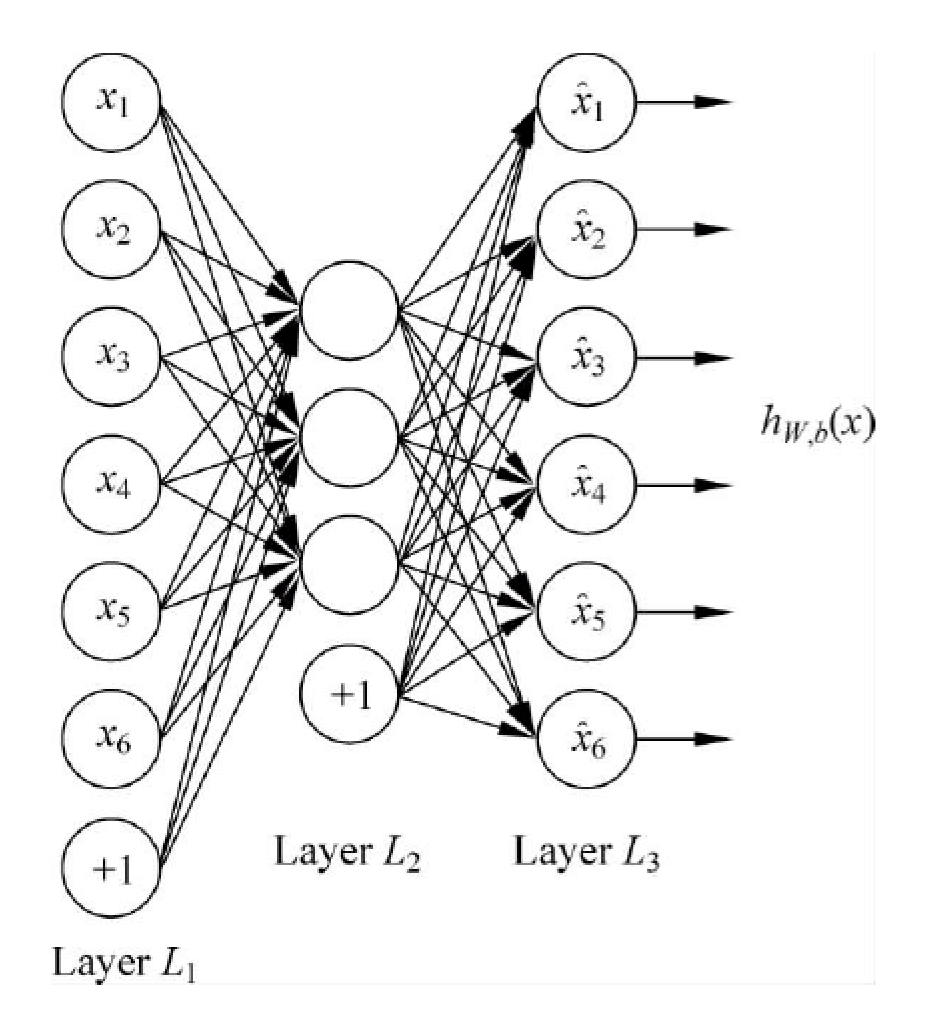


图 7-24　sparse autoencoder 网络结构图

(2) 损失函数。无稀疏约束时网络的损失函数表达式如下

$$J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}J(W,b;\ x^{(i)},y^{(i)})\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(W_{ji}^{(l)})^2$$

$$=\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\parallel h_{W,b}(x^{(i)})-y^{(i)}\parallel^2\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$

稀疏编码是对网络的隐含层的输出有了约束，即隐含层节点输出的平均值应尽量为 0，这样的话，大部分的隐含层节点都处于非激活状态。因此，此时的 sparse autoencoder 损失函数表达式为

$$J_{\text{sparse}}(W,b)=J(W,b)+\beta\sum_{j=1}^{s_2}\text{KL}(\rho\parallel\hat{\rho}_j)$$

后面一项为 KL 距离，其表达式如下

$$\text{KL}(\rho\parallel\hat{\rho}_j)=\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}$$

隐含层节点输出平均值求法如下

$$\hat{\rho}_j=\frac{1}{m}\sum_{i=1}^{m}[a_j^{(2)}(x^{(i)})]$$

其中的参数一般取很小，比如说 0.05，也就是小概率发生事件的概率。这说明要求隐含层的每一个节点的输出均值接近 0.05(其实就是接近 0，因为网络中激活函数为 sigmoid 函数)，这样就达到稀疏的目的了。KL 距离在这里表示的是两个向量之间的差异值。从约

束函数表达式中可以看出，差异越大则“惩罚越大”，因此最终的隐含层节点的输出会接近0.05。

假设有一个固定样本集$\{(x^{(1)},y^{(1)}),\cdots,(x^{(m)},y^{(m)})\}$，它包含$m$个样例。可以用批量梯度下降法来求解神经网络。具体来讲，对于单个样例(x,y)，其代价函数为

$$J(W,b;\ x,y)=\frac{1}{2}\|h_{W,b}(x)-y\|^2$$

这是一个(二分之一的)方差代价函数。给定一个包含m个样例的数据集，可以定义整体代价函数为

$$J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}J(W,b;\ x^{(i)},y^{(i)})\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$

$$=\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h_{W,b}(x^{(i)})-y^{(i)}\|^2\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(il)})^2$$

以上公式中的第一项$J(W,b)$是一个均方差项。第二项是一个规则化项(也叫权重衰减项)，其目的是减小权重的幅度，防止过度拟合。权重衰减参数λ用于控制公式中两项的相对重要性。在此重申一下这两个复杂函数的含义：$J(W,b;\ x,y)$是针对单个样例计算得到的方差代价函数；$J(W,b)$是整体样本代价函数，它包含权重衰减项。

以上的代价函数经常被用于分类和回归问题。在分类问题中，我们用$y=0$或1来代表两种类型的标签，这是因为sigmoid激活函数的值域为[0,1]；如果使用双曲正切型激活函数，那么应该选用-1和$+1$作为标签。对于回归问题，首先要变换输出值域y，以保证其范围为[0,1](同样地，如果使用双曲正切型激活函数，要使输出值域为[-1,1])。

我们的目标是针对参数W和b来求其函数$J(W,b)$的最小值。为了求解神经网络，需要将每一个参数$W_{ij}^{(l)}$和$b_i^{(l)}$初始化为一个很小的、接近零的随机值(比如说，使用正态分布Normal$(0,\epsilon^2)$生成的随机值，其中ϵ设置为0.01)，之后对目标函数使用诸如批量梯度下降法的最优化算法。因为$J(W,b)$是一个非凸函数，梯度下降法很可能会收敛到局部最优解；但是在实际应用中，梯度下降法通常能得到令人满意的结果。最后，需要再次强调的是，要将参数进行随机初始化，而不是全部置为零。如果所有参数都用相同的值作为初始值，那么所有隐藏层单元最终会得到与输入值有关的、相同的函数(也就是说，对于所有i，$W_{ij}^{(1)}$都会取相同的值，那么对于任何输入x都会有：$a_1^{(2)}=a_2^{(2)}=a_3^{(2)}=\cdots$)。随机初始化的目的是使对称失效。

(3) 反向传播算法梯度下降法中每一次迭代都按照如下公式对参数W和b进行更新

$$W_{ij}^{(l)}=W_{ij}^{(l)}-\alpha\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b)$$

$$b_i^{(l)}=b_i^{(l)}-\alpha\frac{\partial}{\partial b_i^{(l)}}J(W,b)$$

其中，α是学习速率。其中关键步骤是计算偏导数。现在来讲一下反向传播算法，它是计算偏导数的一种有效方法。

首先来讲一下如何使用反向传播算法来计算$\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b;\ x,y)$和$\frac{\partial}{\partial b_i^{(l)}}J(W,b;\ x,y)$，这两项是单个样例$(x,y)$的代价函数$J(W,b;\ x,y)$的偏导数。一旦求出该偏导数，就可以推导出整体代价函数$J(W,b)$的偏导数

$$\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b;\ x^{(i)},y^{(i)})\right]+\lambda W_{ij}^{(l)}$$

$$\frac{\partial}{\partial b_i^{(l)}}J(W,b)=\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial b_i^{(l)}}J(W,b,x^{(i)},y^{(i)})$$

以上两行公式稍有不同，第一行比第二行多出一项，是因为权重衰减是作用于 W 而不是 b。反向传播算法的思路如下：给定一个样例(x,y)，首先进行“前向传导”运算，计算出网络中所有的激活值，包括 $h_{W,b}(x)$的输出值。之后，针对第 l 层的每一个节点 i，计算出其“残差”$\delta_i^{(l)}$，该残差表明了该节点对最终输出值的残差产生了多少影响。对于最终的输出节点，可以直接算出网络产生的激活值与实际值之间的差距，我们将这个差距定义为 $\delta_i^{(n_l)}$（第 n_l 层表示输出层）。对于隐藏单元如何处理呢？我们将基于节点（译者注：第 $l+1$ 层节点）残差的加权平均值计算 $\delta_i^{(l)}$，这些节点以 $a_i^{(l)}$ 作为输入。下面将给出反向传导算法的细节。

进行前馈传导计算，利用前向传导公式，得到 $L_2,L_3,\cdots$直到输出层 L_{n_l} 的激活值。

对于第 n_l 层（输出层）的每个输出单元 i，根据以下公式计算残差

$$\delta_i^{(n_l)}=\frac{\partial}{\partial z_i^{(n_l)}}\frac{1}{2}\|y-h_{W,b}(x)\|^2=-(y_i-a_i^{(n_l)})\cdot f'(z_i^{(n_l)})$$

$$\begin{aligned}
\delta_i^{(n_l)}&=\frac{\partial}{\partial z_i^{n_l}}J(W,b;\ x,y)\\
&=\frac{\partial}{\partial z_i^{n_l}}\frac{1}{2}\|y-h_{W,b}(x)\|^2\\
&=\frac{\partial}{\partial z_i^{n_l}}\frac{1}{2}\sum_{j=1}^{S_{n_l}}(y_j-a_j^{(n_l)})^2\\
&=\frac{\partial}{\partial z_i^{(n_l)}}\frac{1}{2}\sum_{j=1}^{S_{n_l}}(y_j-f(z_j^{(n_l)}))^2\\
&=-(y_i-f(z_j^{(n_l)}))\cdot f'(z_i^{(n_l)})\\
&=-(y_i-a_i^{(n_l)})\cdot f'(z_i^{(n_l)})
\end{aligned}$$

对 $l=n_l-1,n_l-2,n_l-3,\cdots,2$ 的各个层，第 l 层的第 i 个节点的残差计算方法如下

$$\delta_i^{(l)}=\left(\sum_{j=1}^{S_{l+1}}W_{ji}^{(l)}\delta_j^{(l+1)}\right)f'(z_i^{(l)})$$

$$\begin{aligned}
\delta_i^{(n_l-1)}&=\frac{\partial}{\partial z_i^{n_l-1}}J(W,b;\ x,y)\\
&=\frac{\partial}{\partial z_i^{n_l-1}}\frac{1}{2}\|y-h_{W,b}(x)\|^2\\
&=\frac{\partial}{\partial z_i^{n_l-1}}\frac{1}{2}\sum_{j=1}^{S_{n_l}}(y_j-a_j^{(n_l)})^2\\
&=\frac{1}{2}\sum_{j=1}^{S_{n_l}}\frac{\partial}{\partial z_i^{n_l-1}}(y_j-a_j^{(n_l)})\\
&=\frac{1}{2}\sum_{j=1}^{S_{n_l}}\frac{\partial}{\partial z_i^{n_l-1}}(y_j-f(z_j^{(n_l)}))^2
\end{aligned}$$

$$= \sum_{j=1}^{S_{n_l}} -(y_j - f(z_j^{(n_l)})) \cdot \frac{\partial}{\partial z_i^{(n_l-1)}} f(z_j^{(n_l)})$$

$$= \sum_{j=1}^{S_{n_l}} -(y_i - f(z_j^{(n_l)})) \cdot f'(z_j^{(n_l)}) \cdot \frac{\partial z_j^{(n_l)}}{\partial z_i^{(n_l-1)}}$$

$$= \sum_{j=1}^{S_{n_l}} \delta_j^{(n_l)} \cdot \frac{\partial z_j^{(n_l)}}{\partial z_i^{n_l-1}}$$

$$= \sum_{j=1}^{S_{n_l}} \left(\delta_j^{(n_l)} \cdot \frac{\partial}{\partial z_i^{n_l-1}} \sum_{k=1}^{S_{n_l-1}} f(z_k^{n_l-1}) \cdot W_{jk}^{n_l-1}\right)$$

$$= \sum_{j=1}^{S_{n_l}} \delta_j^{(n_l)} \cdot W_{ji}^{n_l-1} \cdot f'(z_i^{n_l-1})$$

$$= \left(\sum_{j=1}^{S_{n_l-1}} W_{ji}^{n_l-1} \delta_j^{(n_l)}\right) f'(z_i^{n_l-1})$$

将上式中的 n_l-1 与 n_l 的关系替换为 l 与 $l+1$ 的关系，就可以得到

$$\delta_i^{(l)} = \left(\sum_{j=1}^{S_{n_l+1}} W_{ji}^{(l)} \delta_j^{(l+1)}\right) f'(z_i^l)$$

以上逐次从后向前求导的过程即为“反向传导”的本意所在。

计算我们需要的偏导数，计算方法如下

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = \alpha_j^{(l)} \delta_i^{(l+1)}$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)}$$

最后，用矩阵-向量表示法重写以上算法。使用 · 表示向量乘积运算符(在 MATLAB 或 Octave 里用“.*”表示，也称作阿达马乘积)。若 $a=b \cdot c$，则 $a_i=b_i c_i$。那么，反向传播算法可表示为以下几个步骤。

(1) 进行前馈传导计算，利用前向传导公式，得到 $L_2, L_3, \cdots$ 直到输出层 L_{n_l} 的激活值。

(2) 对输出层(第 n_l 层)，计算

$$\delta^{(n_l)} = -(y - a^{(n_l)}) \cdot f'(z^{(n_l)})$$

(3) 对于 $l=n_l-1, n_l-2, n_l-3, \cdots, 2$ 的各层，计算

$$\delta^{(l)} = ((W^{(l)})^{\mathrm{T}} \delta^{(l+1)}) \cdot f'(z^{(l)})$$

(4) 计算最终需要的偏导数值

$$\nabla_{W^{(l)}} J(W,b;x,y) = \delta^{(l+1)} (a^{(l)})^{\mathrm{T}}$$

$$\nabla_{b^{(l)}} J(W,b;x,y) = \delta^{(l+1)}$$

实现中应注意：在以上的第(2)步和第(3)步中，需要为每一个 i 值计算其 $f'(z_i^{(l)})$。假设 $f(z)$ 是 sigmoid 函数，并且我们已经在前向传导运算中得到了 $a_i^{(l)}$。那么，使用早先推导出的 $f'(z)$ 表达式，就可以计算得到 $f'(z_i^{(l)}) = a_i^{(l)}(1-a_i^{(l)})$。

最后，我们将对梯度下降算法做个全面总结。在下面的伪代码中，$\Delta W^{(l)}$ 是一个与矩阵 $W^{(l)}$ 维度相同的矩阵，$\Delta b^{(l)}$ 是一个与 $b^{(l)}$ 维度相同的向量。注意这里“$\Delta W^{(l)}$”是一个矩阵，而

不是“Δ 与 $W^{(l)}$ 相乘”。下面实现批量梯度下降法中的一次迭代。

对于所有 l，令 $\Delta W^{(l)} := 0$，$\Delta b^{(l)} := 0$（设置为全零矩阵或全零向量）。

(1) 对于 $i=1$ 到 m，使用反向传播算法计算 $\nabla_{W^{(l)}} J(W,b; x,y)$ 和 $\nabla_{b^{(l)}} J(W,b; x,y)$。

(2) 计算 $\Delta W^{(l)} := \Delta W^{(l)} + \nabla_{W^{(l)}} J(W,b; x,y)$。

(3) 计算 $\Delta b^{(l)} := \Delta b^{(l)} + \nabla_{b^{(l)}} J(W,b; x,y)$。

更新权重参数：

$$W^{(l)} = W^{(l)} - \alpha\left[\left(\frac{1}{m}\Delta W^{(l)}\right) + \lambda W^{(l)}\right]$$

$$b^{(l)} = b^{(l)} - \alpha\left[\frac{1}{m}\Delta b^{(l)}\right]$$

现在，可以重复梯度下降法的迭代步骤来减小代价函数 $J(W,b)$ 的值，进而求解神经网络。

7.6　大数据语义分析知识发现

知识发现是人类的主要知识活动之一，当前的知识活动也越来越多地基于网络数据资源环境。在网络资源环境向“语义网”阶段过渡，并已经进入大数据网络时代的时候，知识发现必然面临新的机会和挑战。因此，知识发现也必将是大数据发展和完善的主要动力。大数据时代的知识发现是以结构化数据和非结构化数据为基础，通过数据采集、数据抽取、数据清洗、数据转化、数据加载和数据挖掘等过程，发现可理解、可用的新知识内容，并能在一些领域内加以应用的知识。

从知识发现研究角度来看，基于大数据的知识发现是知识发现的特殊案例。广义的知识发现更加关注于从数据源中发现知识的整个过程，包括数据是如何存储和访问，算法如何自动处理数据并且在大量数据的环境下有效运行，结果如何解释和可视化，以及整个过程中人机交互如何建模和支持。大数据本身是应用新技术、收集、组织和存储的数据资源，是基于互联网发展全球共享超级数据库，如图 7-25 所示。基于大数据的知识活动应当是在遵循数据库知识发现的一般规律时，考虑数据的组织方式、应用工具和技术、资源环境等综合因素。

从知识发现的应用角度来看，知识发现是大数据的一种关键和高层的应用。随着互联网和大数据的快速发展，面向语义网的信息收集、组织、存储和访问的技术和方法也应接不暇，以大数据为基础的数据对象、网络环境、语义关系模型、存取标准（HTTP URI）和网络应用（浏览、搜索等）为知识发现提供了新的路径。如何根据大数据的特点和优势，帮助人们更容易、更准确、更全面、更高效地发现所需要的信息，最终获取准确、实用、即时的知识是大数据知识发现研究的主要方向。

7.6.1　大数据知识发现过程

大数据的知识发现是大数据的高级应用，是在大数据理论、技术、工具和资源环境的基础上的创新性的知识活动。基于大数据的知识发现过程遵循知识发现（Knowledge Discovery in Database，KDD）的一般规律，同时因为技术架构和网络资源环境的变化而有其独特性。其过程如图 7-26 所示。

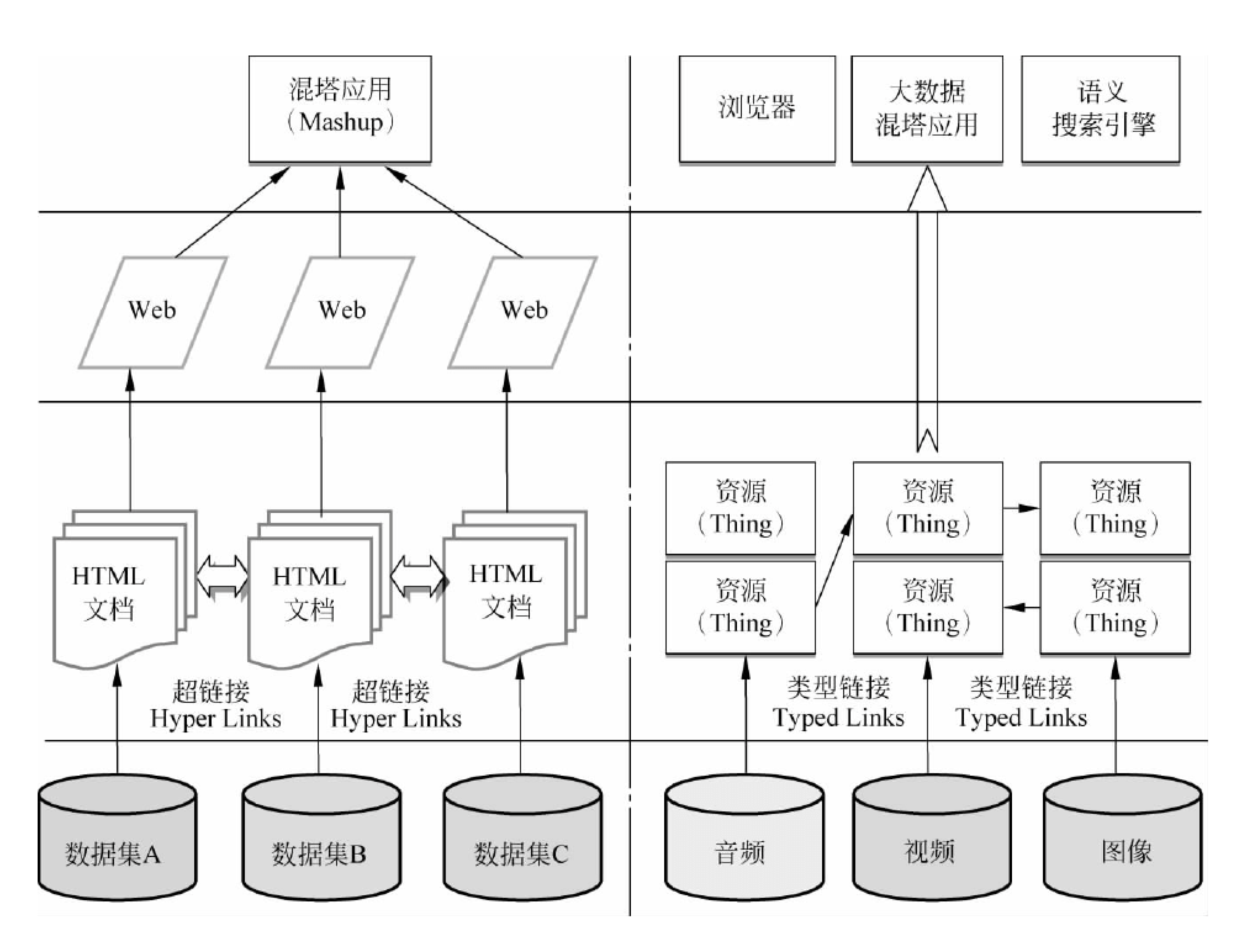


图 7-25 互联网发展全球共享数据源

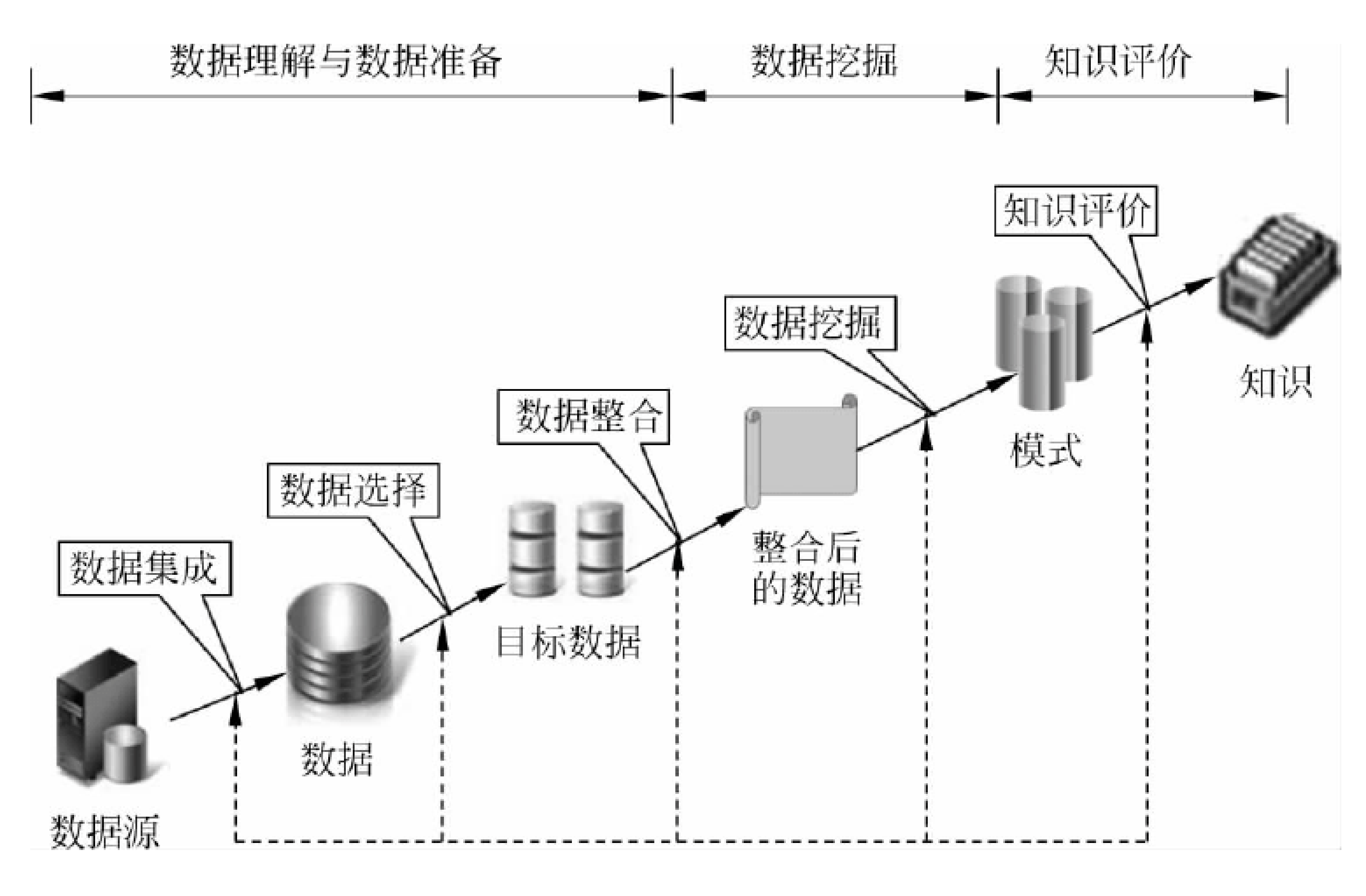


图 7-26 大数据知识发现过程

基于大数据的知识发现过程包括数据集成、数据选择、数据整合和数据挖掘等基本阶段。

1. 数据集成

大数据知识发现的数据集成主要是来源于不同的业务系统数据库、互联网文本、音频、视频等数据在逻辑上或物理上有机地集中，从而提供全面的数据共享。通常采用联邦式、基于中间件模型和数据仓库等方法来构造集成的系统。通过应用间的数据交换从而达到集

成,主要解决数据的分布性和异构性的问题,其前提是被集成应用必须公开数据结构,即必须公开表结构、表间关系、编码的含义等。这一数据准备阶段本身也包含相关数据的链接关系和构建过程,同时为基于大数据的知识发现提供了丰富的语义互连的数据源。

2. 数据选择

大数据网络中数据源数量巨大,并且动态增长,它们来自不同的数据提供者,属于不同的领域,采用不同的本体或者词表术语描述概念,使用不同的访问方式。在这样的海量、异构和动态的数据源上进行知识发现,数据选择是非常重要的步骤。如何能够根据用户查询需求识别和筛选出相关的数据源,同时兼顾完整性、准确性和效率性,是基于大数据知识发现的关键。

3. 数据整合

数据整合基于 ETL 的基本原则进行,其按照统一的规则集成并提高数据的价值,是负责完成数据从数据源向目标数据仓库转化的过程,是实施数据智能管理的重要步骤。主要分为数据抽取、数据清洗、数据转换和数据装载 4 个过程。

数据抽取是指将数据从各种原始的业务系统中读取出来,这是所有工作的前提,在本系统中,当数据收集工作完成后,传递给数据整合模块即实现了数据抽取过程。数据清洗是对数据进行重新审查和校验的过程,目的在于删除重复信息、纠正存在的错误,并保障数据一致性。由于数据来源不同,如物联网、互联网和内部业务系统的数据,因此避免不了有的数据是错误数据,有的数据相互之间有冲突,或者有的数据是无用数据,数据清洗步骤正是要把这些数据处理掉。数据转换是指按照预先设计好的规则将抽取的数据进行转换,使本来异构的数据格式能统一起来。由于网络中大量的数据是非结构化的数据,因此进行适当的数据转换操作,将这些数据统一起来,变成可处理的形式是很有必要的。数据装载是指将转换完的数据按计划增量或全部导入到分布式存储系统中,这是数据整合的最后一步,也即按照一定的规则将整合后的数据传送到分布式存储系统中。

4. 数据挖掘

数据挖掘是知识发现的关键步骤,大数据知识发现除了结构化数据挖掘外,还有非结构化数据挖掘。其中包括文本挖掘和视频挖掘。

文本挖掘是一个从非结构化文本信息中获取用户感兴趣或者有用的模式,对具有丰富语义的文本进行分析从而理解其所包含的内容和意义的过程。其中被普遍认可的文本挖掘定义如下:文本挖掘是指从大量文本数据中抽取事先未知的、可理解的、最终可用的知识的过程,同时运用这些知识更好地组织信息以便将来参考。

我们在日常生活中所能接触到的最普遍的信息存储形式就是文本,研究表明一个企业80%的信息载体是文本文件。文本挖掘是一个多学科领域,涉及信息检索、文本分析、文本分类与聚类、可视化、数据库技术、机器学习和数据挖掘。文本挖掘与三个文本处理技术相关:信息检索(Information Retrieval),文本聚类与分类(Text Classification and Clustering)以及信息抽取(Information Extraction)。信息检索是指信息按一定的方式组织起来,并根据信息用户的需要和查询进行提问,从大量的文本集中找出有关信息的过程和技术。信息检索主要是基于统计的方法来计算理想结果与文本间的相关性。信息检索侧重于发现和抽取文本集中的信息,而没有发掘出新的信息。文本分类是指可以将文本分到预先定义好的

类别里的文本组织技术,而文本聚类主要是依据著名的聚类假设:同类的文档相似度较大,而不同类的文档相似度较小,基于文本所包含信息的相似度将文本聚集。文本分类和聚类只是将文本的内容注释成为相关的关键词列表,也不能导致新的信息被发现。信息抽取是把文本里包含的信息进行结构化处理,变成表格一样的组织形式。输入信息抽取的是原始文本,输出的是固定格式的信息点。信息点从各种各样的文档中被抽取出来,然后以统一的形式集成在一起。

视频挖掘分为广义的视频挖掘和狭义的视频挖掘。广义的视频挖掘,是从大量视频数据中自动提取视频的类别、结构、语义等知识,并且基于这些知识采用传统的数据挖掘方法或者新的视频挖掘方法,发现视频数据或者数据集中的关联、趋势、异常等隐含的、有价值的、可理解的模式。狭义的视频挖掘,不包括为弥补"语义鸿沟"所进行的知识挖掘,仅指从视频数据或数据集中发现内容间的关联、趋势、异常等隐含的、有价值的、可理解的模式。

视频的分析和应用(同样适用于音频、图像)目前处于初级阶段,同时在挖掘技术上还面临着诸多挑战,如特征的有效提取和快速检索,针对海量文件的分布式并行挖掘算法改造等。由于不同领域的特点和应用目的不尽相同,需要针对具体关联领域研究开发新的挖掘方法。

5. 大数据可视化

如今,数据生产的速度远远超过了数据消化的速度,数据类型也不仅仅是结构化的,这些数据属性的变化为数据交互和展示带来了新的挑战,如实时数据可视化分析报告、交互式和动态图形或报告的可视化及海量数据的可视化等。传统的结构化数据统计分析方法和可视化展示方式就很难满足快速地掌握数据内部规律和变化趋势的需要。

非结构的可视化技术并不一定能准确给出计算结果,但其价值在于能够支持快速地找出这些结果。数据的内部规律和变化趋势不是由数字而是通过可视化对象来描述的。

如图 7-27 和图 7-28 所示,利用探索驱动、Fail-fast 的方法分析非结构化数据,以便于更好地理解业务问题。



图 7-27 标签云-海量文本高频展示



图 7-28 数据占比信息快速获取

7.6.2 大数据知识发现技术框架

实现大数据知识发现,运用大数据技术从数据源数据集成到知识获取和展示应用是一个综合系统化的过程,主要技术应用归纳如下。

(1) 数据采集技术(条码、二维码、传感器、RFID、GIS、GPS、生物识别、移动技术等);

(2) 数据整合与管理(去噪、排重;ETL/ELT;元数据管理;主数据管理;数据质量管理);

(3) 存储技术(分布式文件;行、列存储;NewSQL DB;NoSQL DB;云存储;一体机);

(4) 语义技术(量化数据、文本、视频、声音和图像;用于数据挖掘、机器学习和知识管理);

(5) 计算技术(网格计算:分布式计算;并行计算、内存计算、流线批处理、实时交互计算、流计算;计算资源虚拟化等);

(6) 分析技术(ROLAP、MOLAP、Hive、Impala、Shark 等);

(7) 数据挖掘(R 语言、Mahout、算法库 & 行业模型库);

(8) 数据可视化(设备多样化;文本、地图、仪表板技术;立体可视化、流量可视化、可视化交互等);

(9) 数据安全(用户、认证、授权、审计;数据脱敏;合规和企业内控);

(10) 运维管理工具(资源管理、调度管理、部署管理、流程管理、监控;PasS 管理等)。

7.6.3 大数据知识发现专家系统

专家系统是一个智能计算机程序系统,其内部具有大量专家水平的某个领域知识与经验,能够利用人类专家的知识和解决问题的方法来解决该领域的问题。也就是说,专家系统是一个具有大量专门知识与经验的程序系统,它应用人工智能技术,根据某个领域一个或多个人类专家提供的知识和经验进行推理和判断,模拟人类专家的决策过程,以解决那些需要专家决定的复杂问题。

当前的研究涉及有关专家系统设计的各种问题。这些系统是在某个领域的专家(他可能无法明确表达他的全部知识)与系统设计者之间经过艰苦的反复交换意见之后建立起来的。在已经建立的专家咨询系统中,有能够诊断疾病的(包括中医诊断智能机),估计潜在石油等矿藏的,研究复杂有机化合物结构的以及提供使用其他计算机系统的参考意见等。发展专家系统的关键是表达和运用专家知识,即来自人类专家的并已被证明对解决有关领域内的典型问题是有用的事实和过程。专家系统和传统的计算机程序最本质的不同之处在于专家系统所要解决的问题一般没有算法解,并且经常要在不完全、不精确或不确定的信息基础上做出结论。

专家系统可以解决的问题一般包括解释、预测、诊断、设计、规划、监视、修理、指导和控制等。高性能的专家系统也已经从学术研究开始进入实际应用研究。随着人工智能整体水平的提高,专家系统也获得发展。基于大数据支持处理下的专家系统有分布式专家系统和协同式专家系统等。在新一代专家系统中,不但采用基于规则的方法,而且采用基于模型的原理。如图 7-29 所示,知识发现专家系统模型图,主要包括 KDD 集成、协调器、人机交互知

识获取和KDK等部分组成。

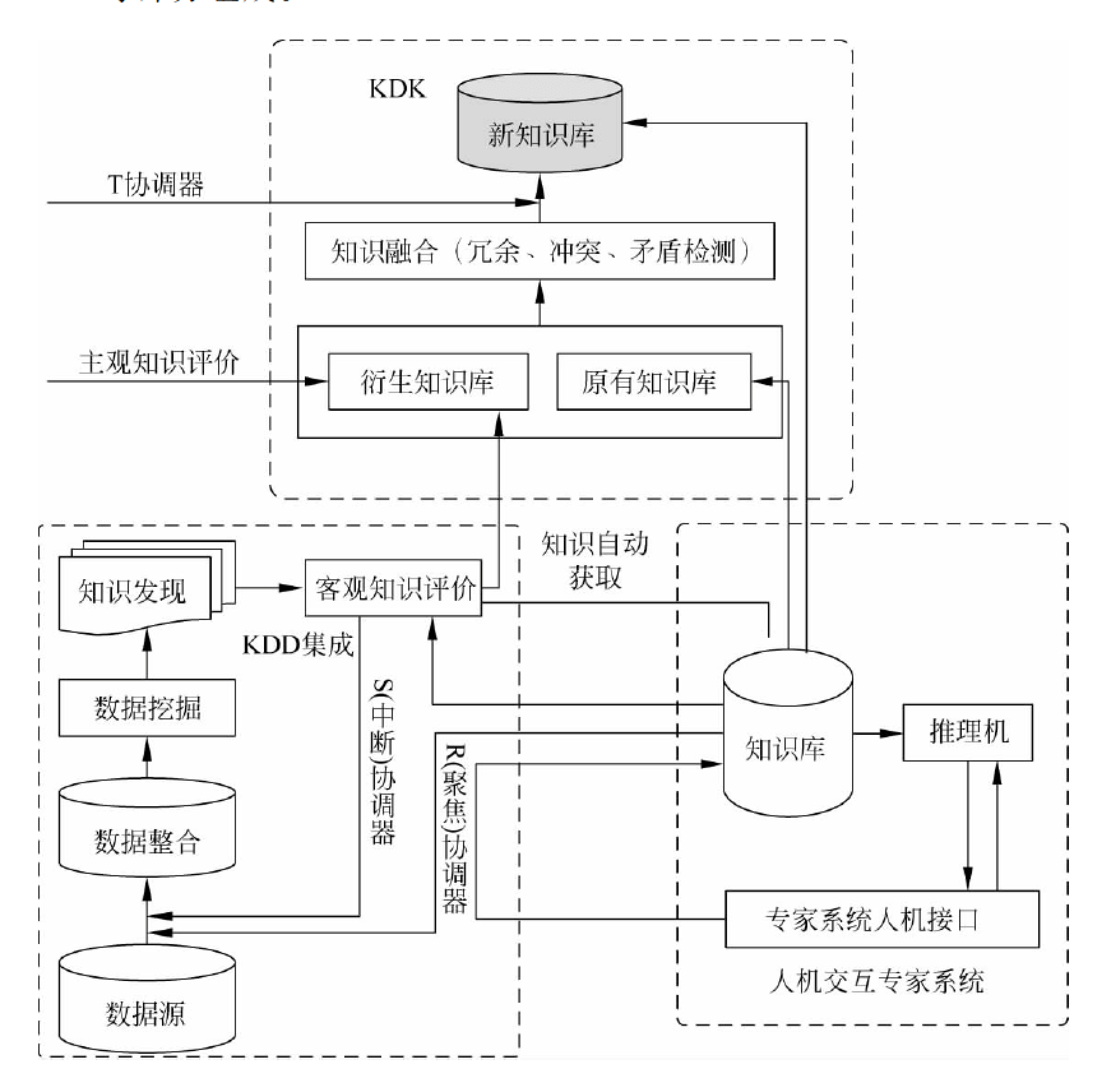


图7-29 知识发现专家系统模型图

1. KDD集成

在知识发现系统中，KDD算法在专家系统知识库的引导下，由R协调器对基本数据库进行聚焦，由S协调器中止挖掘进程。

R(聚焦)协调器根据专家系统知识库对数据源进行聚焦，从而得到挖掘数据表。R协调器实现以下三种聚焦。

(1) 知识类型聚焦。通过对专家系统知识库中相应知识体的分析，R协调器可以建议用户应该挖掘什么类型的知识，比如分类规则还是回归式。

(2) 数据聚焦。R协调器抽取出和专家系统知识库中相应知识体相关的属性字段，组成挖掘数据表。那些知识库中没有出现的字段，将不会在挖掘数据表中。这使得挖掘数据表更具有针对性。

(3) 知识形式聚焦。这主要针对关联规则挖掘和分类规则挖掘而言。通过一种规则模板的形式，R协调器可以指定要挖掘的规则的具体形式，从而挖掘更感兴趣的知识。规则模板可以由R协调器分析知识库得到。

如果客观知识评价发现挖掘到的知识在知识库中已经存在，则S(中断)协调器发挥作用，中断下一步的知识库精化过程，从而避免耗时去做无谓的知识融合。

2. 协调器

R 聚焦协调器根据领域专家或用户提供的元知识、专家系统知识库的一致性和有效性检验结果、专家系统推理运行失败的断点，对数据源进行聚焦，从而得到挖掘数据表。例如，元知识“发现平均气温和日照时数与雌虫密度之间的关系”，则 R 协调器将从虫害发生数据库中抽取平均气温、日照时数、雌虫密度三个属性的数据，生成挖掘表。元知识类似于一个指定架构的挖掘任务。而对于“平均气温和雌虫密度之间缺乏关联”的知识库检验结果，R 协调器将生成由平均气温和雌虫密度构成的挖掘数据表。对于“知识对象‘病害诊断’推理失败”的推理运行失败断点，R 协调器将抽取所有在知识对象“病害诊断”中出现的属性数据，构成挖掘表。R 协调器使得数据挖掘过程更具有针对性。

对数据挖掘得到的规则性知识首先进行知识评价。如果知识已经存在于知识库中(随着系统运行时间增加，这种情况最终会发生)，则 S 中断协调器发生作用，中断后续的知识融合过程，回到数据挖掘的开始，此时可能需要选择新的数据进行挖掘。

3. 人机交互专家系统

专家系统是人工智能应用研究最活跃和最广泛的课题之一。自从 1965 年第一个专家系统 DENDRAL 在美国斯坦福大学诞生以来，仅经过二十多年的研究发展，到 20 世纪 80 年代中期，各种专家系统就已遍布各个领域，取得很大成功。

最初的专家系统定义是：专家系统是一个具有大量的专门知识与经验的程序系统，它应用人工智能技术和计算机技术，根据某个领域一个或多个专家提供的知识和经验，进行推理和判断，模拟人类专家的决策过程，以便解决那些需要人类专家处理的复杂问题。简言之，专家系统是一种模拟人类专家解决领域问题的计算机程序系统。

对于 20 世纪的专家系统研究，可以归纳出以下几点。

(1) 专家系统最主要的部分是知识库和推理机。知识库用于存放领域专家的知识，包括事实、可行操作和规则等推理机用于记忆采用的规则和控制策略的程序，使专家系统能够以逻辑方式协调工作。推理机根据知识进行推理和导出结论，而非简单的搜索。

(2) 知识库是专家系统发展出的很重要的思想，它不仅促进了人工智能的发展，而且对整个计算机科学的发展影响甚大。

(3) 建立知识库涉及知识获取和知识表示。最初的知识获取指知识工程师从领域专家那里获得知识，知识表示则用计算机能够理解的形式表示和存储这些知识。

(4) 推理机模拟人类专家解决问题的思路。这种方式对于结构化问题求解效果良好，而对于非结构化的问题则往往无能为力。这种“模拟”的思路使得专家系统得以在初期蓬勃发展，但也阻碍了专家系统的进一步发展。

近来一些研究者认为，人工智能是对各种定性模型(物理的、感知的、认识的和社会的系统模型)的获得、表达及使用的计算方法进行研究的学问，从这个意义上说，一个专家系统中的知识库应该是由各种模型综合而成的，而这些模型又往往是定性的模型。由于模型的建立与知识密切相关，所以有关模型的获得、表示及使用自然地包括知识获取、知识表示和知识使用。以这样的观点来看待专家系统的设计，可以认为一个专家系统是由一些原理与运行方式不同的模型综合而成。最近一些研究认为，发展专家系统不仅要采用各种定性模型，而且要运用人工智能和计算机技术的一些新思想与新技术，如分布式、协同式和学习机

制等。

正如专家系统的先驱费根鲍姆(Feigenbaum)所说：专家系统的力量是从它处理的知识中产生的，而不是从某种形式主义及其使用的参考模式中产生的。专家系统的水平完全依赖于它所拥有的知识，而知识获取历来是开发专家系统的一个瓶颈。在早期，知识获取被视为一个从人类知识到特定知识库的转换过程。这种转换基于知识已经清晰存在，只需要搜集并加以表示的假设。这些知识一般是通过对特定领域专家进行咨询得到，并被表示为产生式规则。这种基于知识转换的知识获取具有以下缺点：知识转换困难，知识工程师和领域专家之间存在隔阂，从而使得领域专家的知识很难被规范地表示出来；转换前提有时难以成立，知识也许存在，但并非总是为领域专家所知，而且领域专家有时依靠"经验"行事；表示形式有限，知识转换一般将知识表示为产生式规则，对于那些存在于数据中的隐性知识难以转换和表示。

近来一些研究认为，知识库系统的开发过程更应被视为一个建模过程，建立知识库系统意味着建立一个具有专家能力的求解问题的计算机模型。其本质不在于模拟专家解决问题的过程，而在于建立能达到相似效果的知识模型。知识获取不再被视为对知识进行转换，而是成为建模过程的一部分。传统知识转换得到的规则仅仅是知识模型的一种。

自熊范纶等提出农业专家系统以来，经过近二十年的发展，已经得到广泛应用，但仍然依赖于农业专家提供知识，而耗费巨资普查得到的作物苗情、土情、肥情、病虫害、气象等大量数据资料，基本作为文件存档等。将发现的知识与专家系统知识库有效融合，促进专家系统的自动知识获取和知识精化，具有重要的理论和实际意义。

经过数十年的发展，数据挖掘技术已经逐渐走向实用化阶段。如何减轻那些缺乏专业知识的最终用户由于使用数据挖掘带来的技术上的压力，减轻操作负担，使得他们可以将注意力集中在使用数据挖掘的真正目的上——获取和使用知识，是一个迫切需要解决的问题。另一方面，专家系统发展到今天，已获得了广泛的应用。知识决定着专家系统的能力，而知识获取却仍然是建造专家系统的瓶颈。知识发现给专家系统的自动知识获取带来希望。二者之间有效的集成最终决定知识发现的实用性。知识发现自身也强调，所发现的知识的价值存在于它的适当使用中。将知识发现应用到专家系统中，将大大改善专家系统的知识获取能力，从而提高专家系统的决策能力，在促进专家系统深入发展的同时，也将促进知识发现的更广阔应用。

4. KDK

KDK 进行衍生知识库(存放发现的知识)与专家系统知识库的合成和提炼，并可启动 T 协调器，与领域专家进行交互，生成扩展知识库，利用动态变化后的真实数据库或新数据库，在下一个抽象级上进行 KD(D&K)的知识精化。由此循环，实现知识库中知识的不断精化与提升。

挖掘出的知识经过知识评价后将被存储在中间知识库中。中间知识库与专家系统原有知识库进行知识融合，包括冗余、冲突、矛盾检测。T 协调器在这里实际代表着应用领域专家，尽管他们不一定知道数据挖掘和知识融合的具体过程，但对于那些相互冲突、相互矛盾的知识，他们也许想自己决定如何处理。基于超图的知识表示技术被用来表示知识，从而发现知识的冗余、冲突和矛盾。

对于智能系统来说，知识库不应该是一成不变的。首先可能存在适用性的问题。某个

知识库可能只是反映了某个地区的情况。这在农业领域是非常普遍的。以植保为例,不同的地区有着不同的病虫害,同样的病虫害有着不同的灾害发生规则。其次当数据不断积累时,知识库可能变得不再有效。原有的知识需要修改,以适应数据增长。在知识库精化过程中,由数据挖掘得到的知识与原有智能系统知识库融合,从而使得知识库能够得到完善和精化。我们采用专家系统开发出施肥知识发现系统和植保知识发现系统。施肥知识发现系统可以应用于科学施肥。用户只需提供自己的数据库和知识库,就可进行诸如土壤肥力评估、施肥量确定、目标产量确定等方面的知识发现和知识库精化。植保知识发现系统与施肥知识发现系统具有相同的系统结构,不同的是该系统结合植保领域知识,应用于病虫害诊断等植保领域。这两个系统易学易用,操作傻瓜化,实际运用证明可以有效发现和精化知识。尤其是植保知识发现系统,发现了虫害不同代数发生之间的序贯关系,这是过去植保专家系统知识库中所没有的知识。

知识发现系统面向应用领域的一般用户。用户操作知识发现系统的过程就像是在解决他所熟悉的领域问题;同时系统自动运行的特性使得用户可以集中精力分析处理数据挖掘得到的知识,大大减轻了用户操作负担,提高了数据挖掘技术的实用性。发现的知识最终通过 KDK 过程融合到专家系统知识库中,实现知识精化。

7.6.4 企业大数据知识管理框架

企业从数据中提取知识后,还需要重视知识积累和知识提炼,只有建立完整的知识管理体系和流程,才能够实现对大数据的充分利用。知识管理是网络新经济时代的新兴管理思潮与管理方法,著名管理大师彼得·德鲁克在 1965 年即预言:“知识将取代土地、劳动、资本与机器设备,成为最重要的生产因素。”

知识管理系统,即根据知识管理理论、客户实际状况,完成对组织中大量有价值的方案、策划、成果、经验等知识进行分类存储和管理,积累知识资产,避免知识资产流失,促进知识的学习、共享、培训、再利用和创新,有效降低组织运营成本,强化其核心竞争力的软件系统。知识管理系统作为知识管理过程中最主要的生产、应用、分析系统,从工具性的角度提供了知识的创新、审核、发布、使用、交互、共享、推送、评价、考核、分析、分拣等具体的功能。图 7-30 是现代企业新一代知识管理系统架构图,它汇总了内、外部不同来源的知识,并提供了知识挖掘、知识地图、专家网络等多种信息处理、知识抽取、知识管理等功能支持的知识管理系统框架图。

大数据改变了企业数据利用和知识管理的现状,并进一步改变了传统企业主要依靠经验的企业决策方式,使得企业经营者可以借助海量数据和先进的数据分析手段,得到更加有据可依的经营建议,而这也将对企业的决策模式带来影响。对于企业高层管理者来说,以往的决策过程主要依赖个人经验和直觉的判断,或者简单数据分析,而立足于充分数据分析的决策模式将帮助企业管理者以更加科学的方式完成决策过程,提高决策的准确度。在以往的决策过程中,企业的一般员工因为对企业全貌缺乏把握,难以提出对企业决策的全局性建议,也就无法参与企业的核心决策过程。但是在大数据时代,企业数据中心的建立和企业知识管理流程的科学化、规范化和公开化使得普通员工也能够获得充足的企业决策信息,使得更多的普通员工能够了解企业的整体动向并对企业变革或改进提出合理化建议。而这些变化无疑也将改变原有的企业运营模式,使得企业的组织架构、决策模式进一步向扁平化发

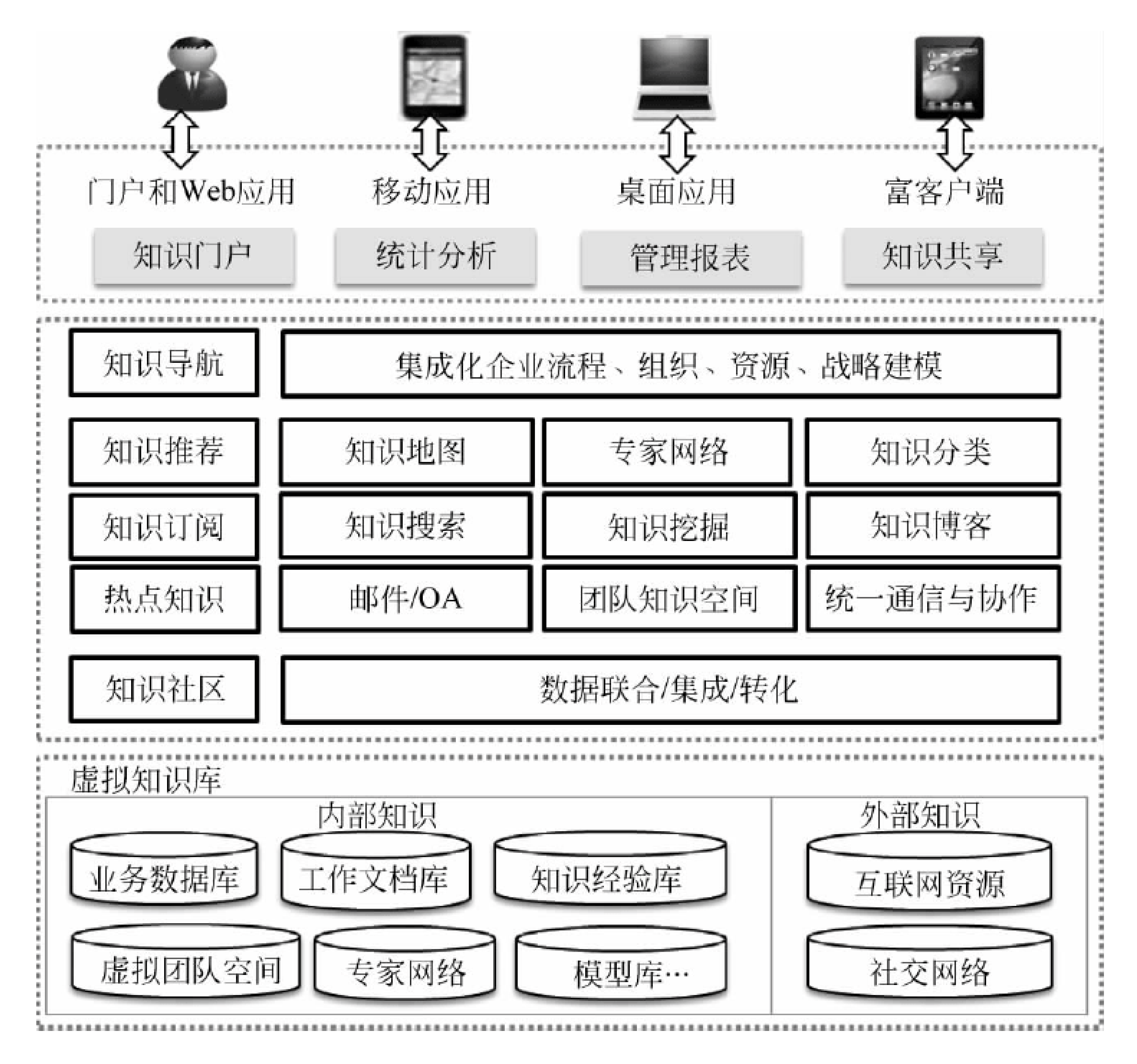


图 7-30 现代企业新一代知识架构图

展，企业管理者将和普通员工一起完成决策过程。大数据影响下的决策环境更加复杂，决策时效性更高，而传统的集权式决策模式也将向着分散式的决策方式进行转变，企业决策以经营数据为依据，企业的决策过程更加趋向于扁平化的调整。

企业数据中心建立后，各部门之间的数据和信息互连互通，变得更加透明化，来自设计、生产、销售、供应链、服务等不同部门的数据能够被有效整合，形成完整、清晰、精细的产品和用户信息流。因此，企业各部门之间的合作也将变得更加简单便捷，有助于企业部门之间边界的模糊化，极大地提升企业价值创造和业务流程效率。

当然，实现企业大数据背景下基于知识管理的企业决策系统，要通过企业信息化系统的顶层设计和分布实施，进行统一思想和树立大数据思维模式全员参与的意识，并且进行业务系统知识培训和数据收集、整理、分析、决策过程规划指导，力求通过建立企业大数据中心和知识管理系统，使企业能够更加关注数据、分析数据、利用知识系统科学决策、创造全员参与的整体工作氛围，促进企业有效、快速、健康可持续发展。

7.7 大数据分析处理平台

7.7.1 结构化大数据处理架构

结构化大数据处理，涉及企业处理的多个环节，从捕获、存储、计算到分析挖掘，可以作为性能提升和解决方案单独部署。主要功能包括数据集成、实时数据同步、数据仓库、分析引擎和数据挖掘等。

1. 数据集成

数据集成(DI)主要涵盖传统 ETL 能力,提供了丰富的数据处理、转换功能组件,同时可集成 CDC 工具、主数据管理产品(MDM),为企业提供一个全面的数据集成处理解决方案。CDC 工具除了可集成到数据集成产品中提供实时能力之外,也可以单独部署,满足企业实时数据同步、灾备等需要。

数据集成主要使用场景为从业务数据源到数据仓库系统的 ETL 过程。数据集成成品通过支持 Web Service 可与企业服务总线产品 ESB 进行数据交互,通过与主数据管理产品集成可降低数据仓库应用的复杂度,同时也可以为主数据管理提供支持。一般来说,数据集成可适用于任何从源数据到目标数据的处理转换。其集成技术架构如图 7-31 所示。

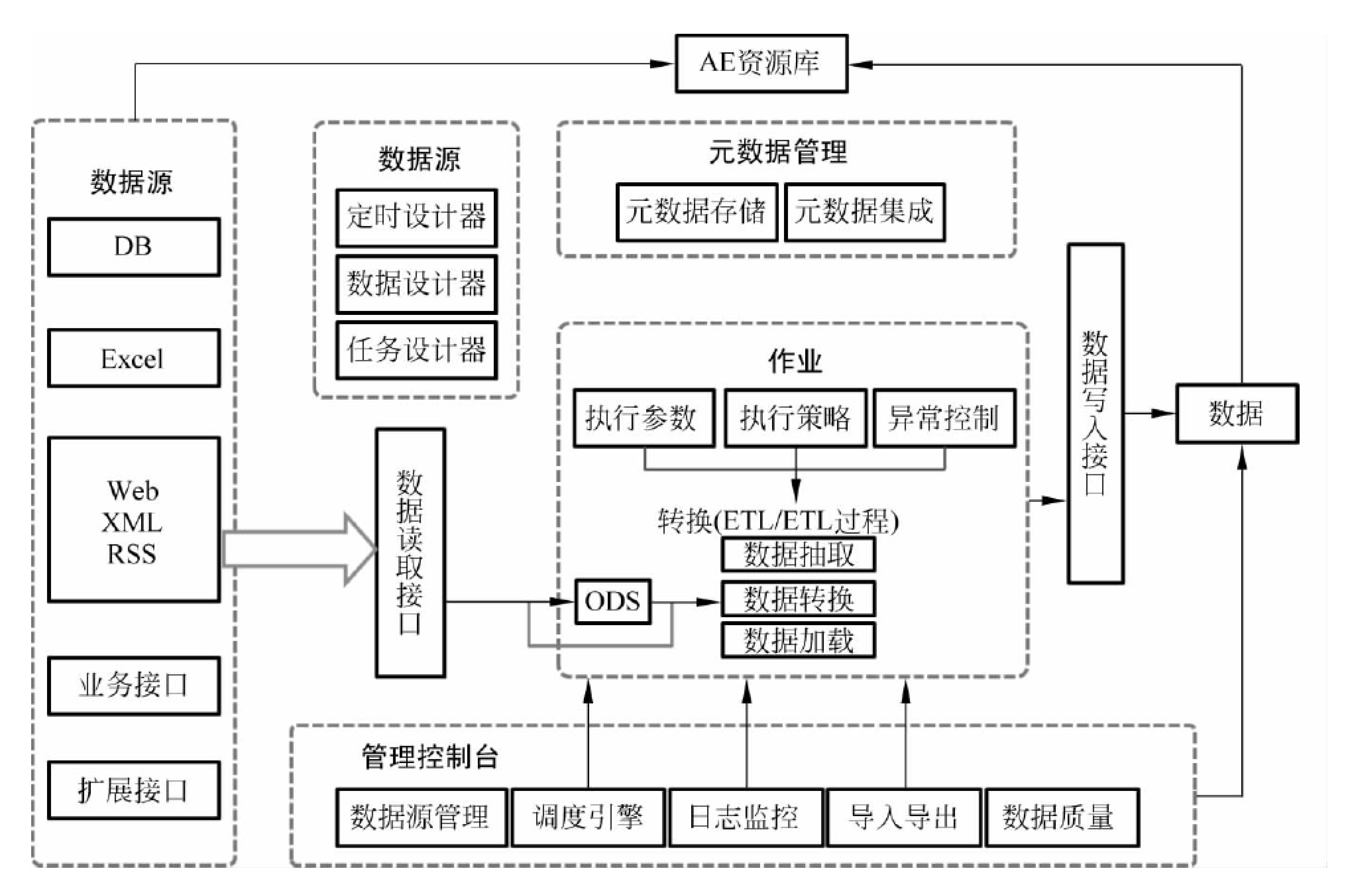


图 7-31　大数据集成技术结构图

2. CDC 工具

CDC(Change Data Capture)工具是基于日志分析和消息中间件技术,内部具有高缓存、高并发的架构,实现了高性能的增量式数据复制和灵活的部署模型。CDC 工具能够提供面向数据仓库的高效数据加载以及异构系统间数据的实时同步。其工作原理简单描述如图 7-32 所示。

其技术特点如下。

(1) 基于数据库日志的增量获取技术,减少对生产库性能的影响;

(2) 采用消息中间件技术,支持灵活部署,并具备异常处理机制,稳定可靠;

(3) 提供完善的管理和监控工具;

(4) 可支持 1000 个在线用户产生的业务数据,集成延迟小于 3s。

3. 分析引擎

OLAP 引擎的核心作用是接收前端工具或应用的多维分析操作发送的请求,基于数据

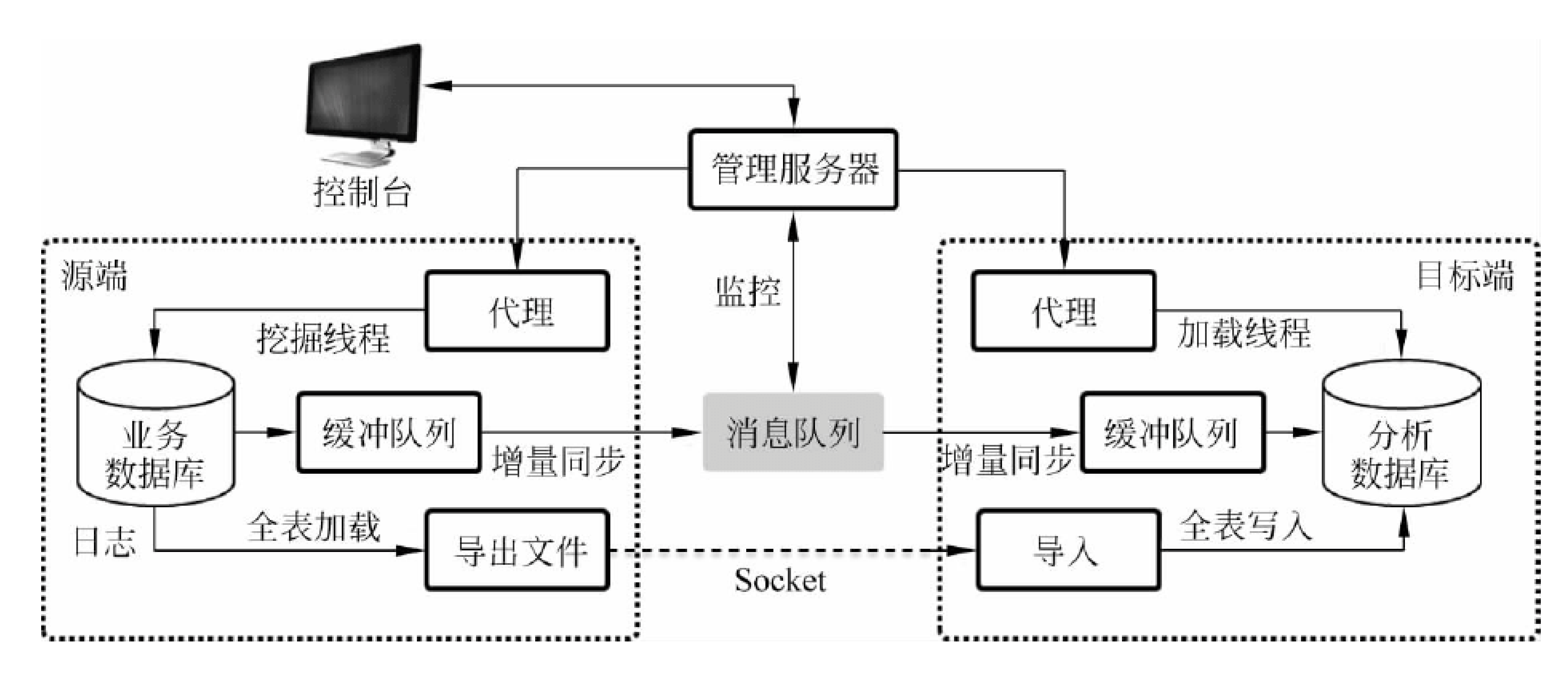


图 7-32　CDC 工作原理图

仓库进行数据查询、聚合，最后返回给前端工具或应用处理结果。OLAP 引擎是基于 HOLAP(Hybrid OLAP)技术架构的多维分析引擎，分析数据存放在关系数据库中，聚合结果存于高速缓冲中，其技术架构如图 7-33 所示。

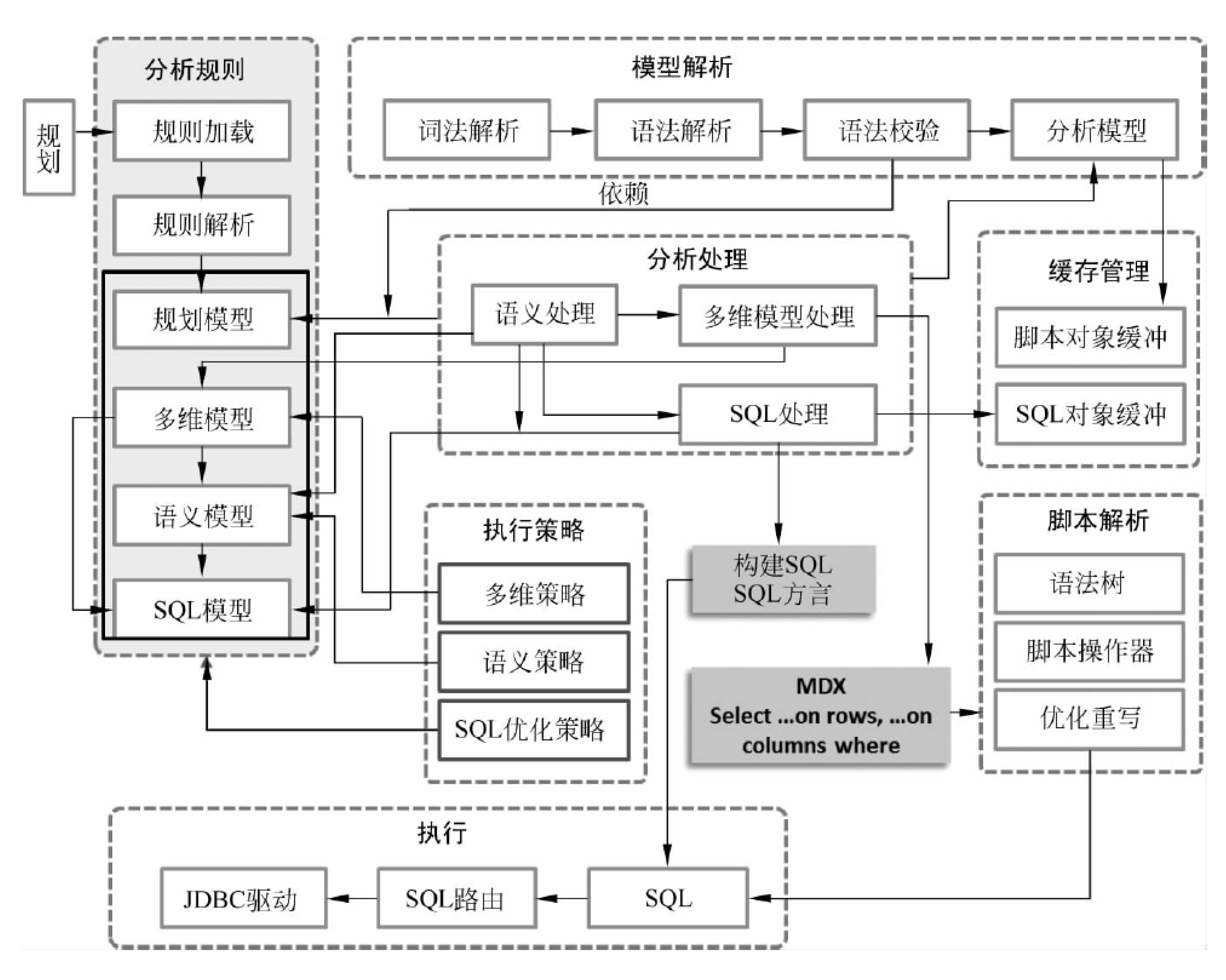


图 7-33　OLAP 分析引擎

4. 数据挖掘

E-Miner 架构如图 7-34 所示，其提供了种类丰富的数据处理、挖掘预测、可视化等组件。通过探索和挖掘企业运营数据中潜在的各种关系、规律和趋势，抽象提炼数据模型，并

通过模型的可视化、模型的发布、模型的预警等功能为用户快速进行运营策略的制定和调整提供支撑。E-Miner 遵循数据挖掘标准 CRISP-DM，是为用户提供基于模型的全生命周期管理的挖掘项目实施方法论。

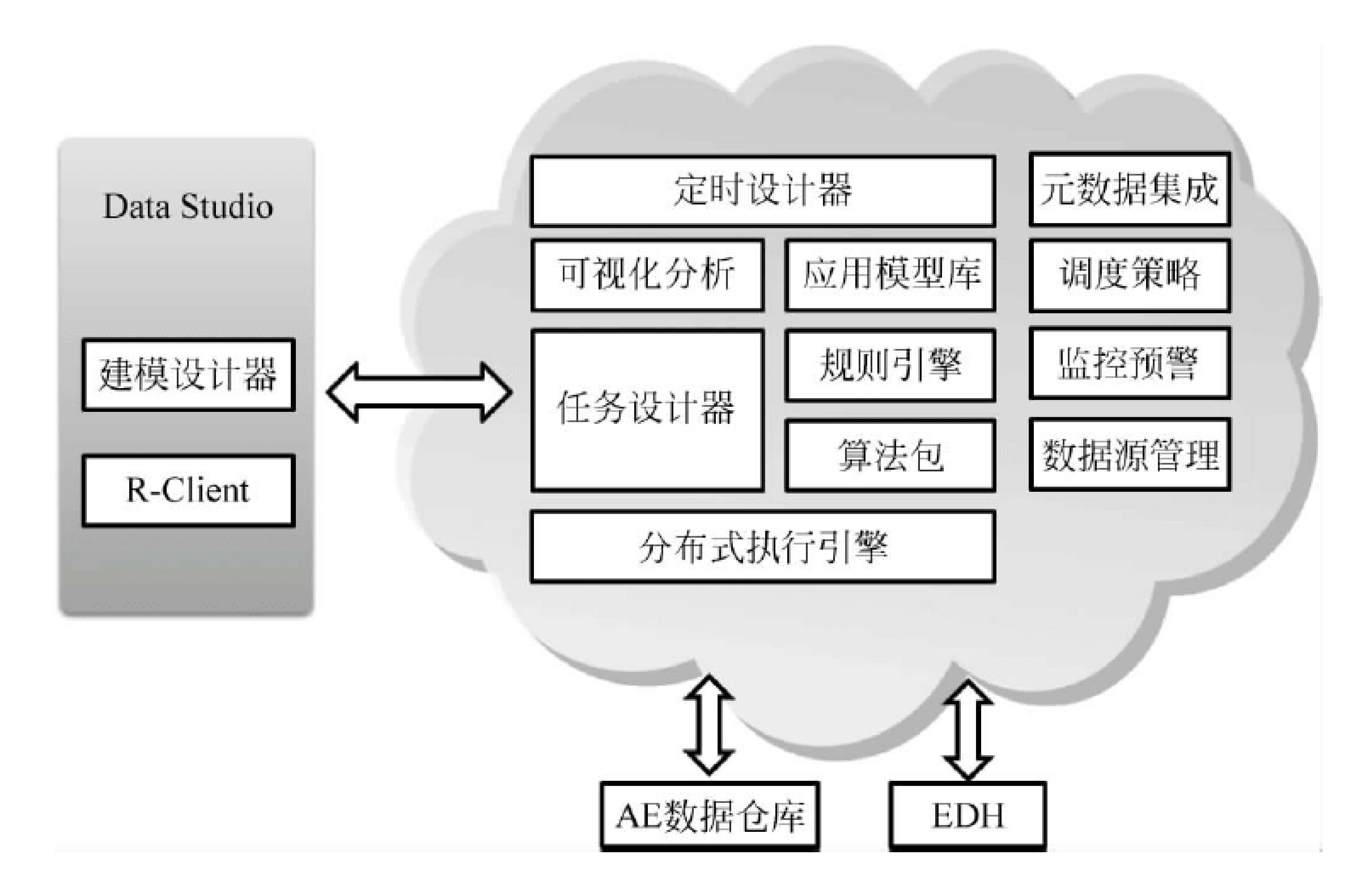


图 7-34　E-Miner 架构图

(1) 数据处理组件。E-Miner 支持数据预处理、统计分析、挖掘和预测算法，同时采用优化的算法迭代机制和基于分布式并行技术的高速数据处理引擎，极大地提高了算法执行效率。具体算法分类如表 7-18 所示。

表 7-18　E-Miner 支持算法归类

分类分析	决策树、线性回归、逻辑回归、贝叶斯、神经网络分类、支持向量机分类等
聚类分析	*k*-means、基于 *k*-means 的层次聚类、分类估计聚类、两阶段聚类等
关联分析	购物篮分析、属性关联分析、序列模式分析
时间序列	滑动平均值、指数平滑、自回归差分滑动平均-ARIMA、趋势估计
预处理	抽样、划分、正规化
统计分析	描述性统计(归纳、列表)、数据探查(拟合、离散化、估计、因子分析)、异常检测、层次聚类等

(2) 集成 R。R 是一门用于统计分析和数据可视化的开源编程语言和软件框架，比商用的挖掘工具(SPSS+SAS)支持更多的算法包，其主要涵盖了概率分析、机器学习、时间序列分析等。

E-Miner 与 R 深度集成，提供 R 语言的开发调试环境，同时可以将自定义的 R 包发布成挖掘算法组件，加入到算法库中。

(3) 数据可视化。E-Miner 提供了折线图、柱状图、散点图、K 线图、饼图、雷达图、地图、和弦图、力导向布局图、仪表盘以及漏斗图来展现数据，同时支持任意维度的堆积和多图表混合展现，利用用户对数据和模型的观察和理解。同时挖掘平台提供了可定制的可视化接口，可以根据数据分析的要求灵活地控制可视化的效果。

7.7.2　非结构化大数据处理架构

EDH(Enterprise Distribution for Hadoop)企业级非结构性大数据处理主要处理大量

的非结构化或半结构化类型数据，也适用于超大规模的结构化数据处理分析，如图 7-35 所示。

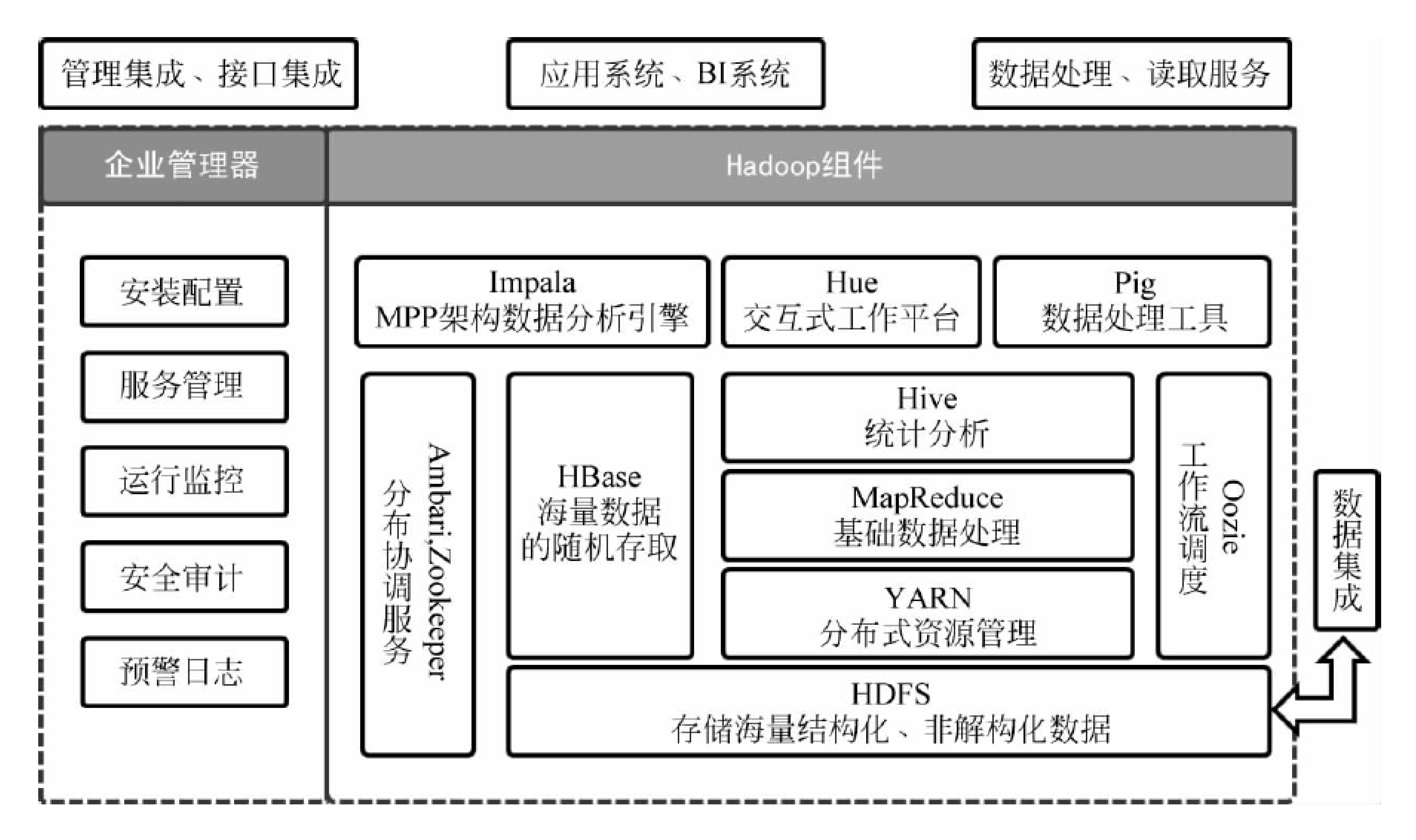


图 7-35　非结构化大数据处理框架

Hadoop 是开源的分布式架构，能够让用户在不了解分布式底层运行细节的情况下，对数据进行分布式处理，能够充分利用分布式集群的高效计算和存储能力。EDH 在开源社区软件的基础上，包含 Hadoop 大部分的主流组件，并且对这些组件在安全性、管理、性能、高可用性等方面进行了优化。同时整合数据集成工具，增强了其企业级应用特性，让企业可以更快、更准、更稳地从各类繁杂无序的海量数据中洞察商机。

EDH 主要用于解决企业的以下需求问题。

(1) 快速整合，存储，集中管理不同类型的海量数据；

(2) 提供批量和实时数据处理服务；

(3) 为构建企业级数据仓库提供大数据平台支撑；

(4) 结合商务智能和数据挖掘可视化产品，提供数据分析服务；

(5) 提供平台中服务组件的管理和系统运行监控。

1. 企业大数据处理架构

EDH 中主要包含 Hadoop 数据处理组件和企业管理器两大部分。其中，Hadoop 组件涵盖了批量处理和实时查询两种处理服务。在数据处理组件部分通过性能优化大幅提升其处理效率和处理能力；企业管理器中包含运维管理、监控服务、安全等多方面内容。具体如图 7-36 所示。

EDH 的技术优势主要体现在：与 Hadoop 标准兼容，支持分布式并行计算、无单点故障；支持结构化、半结构化和非结构化数据的存储、管理和探查；支持大规模集群节点的监控和状态预警，提供自动故障恢复机制保证系统的高可用；兼容 X86 硬件体系架构，可以在廉价服务器上部署，降低总体成本；支持实时的结构化和非结构化数据访问；友好的集群管理界面，实现对各个组件的方便管理；能够模块化地交付多样化、个性化的业务功能，可将大数据系统与现有 IT 系统进行无缝整合，可以与传统商业智能产品和数据挖掘产品无

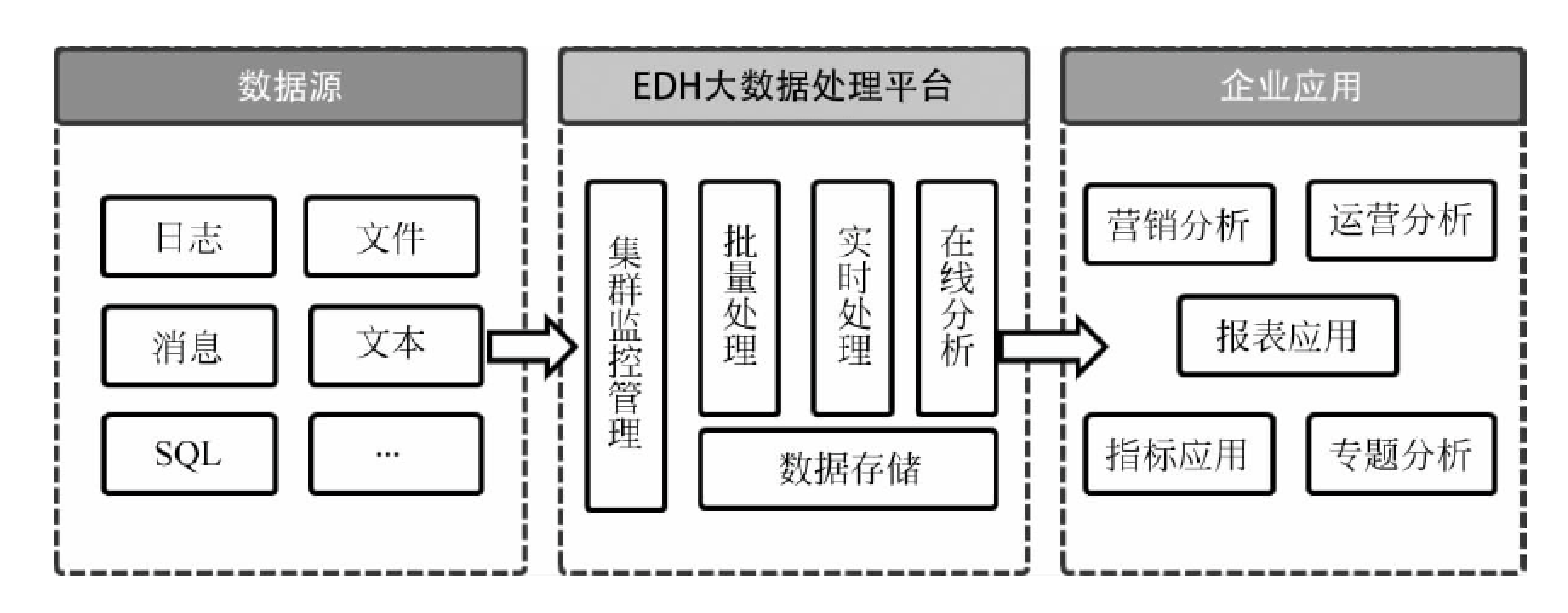


图 7-36 EDH 产品架构

缝集成。

2. 大数据处理解决方案

基于 EDH 数据处理组件与数据集成和 CDC 工具整合，可以为企业在处理大数据中批量处理和实时处理两种场景提供对应解决方案，如图 7-37 和图 7-38 所示。批量处理部分利用 MapReduce、分布式计算引擎技术，具有高并发、大容量特性可以大幅度提高数据处理效率。实时数据处理部分结合 HBase 的高速存取和 Impala 的高效处理能力与 CDC 工具衔接，完成数据的实时采集和分析。

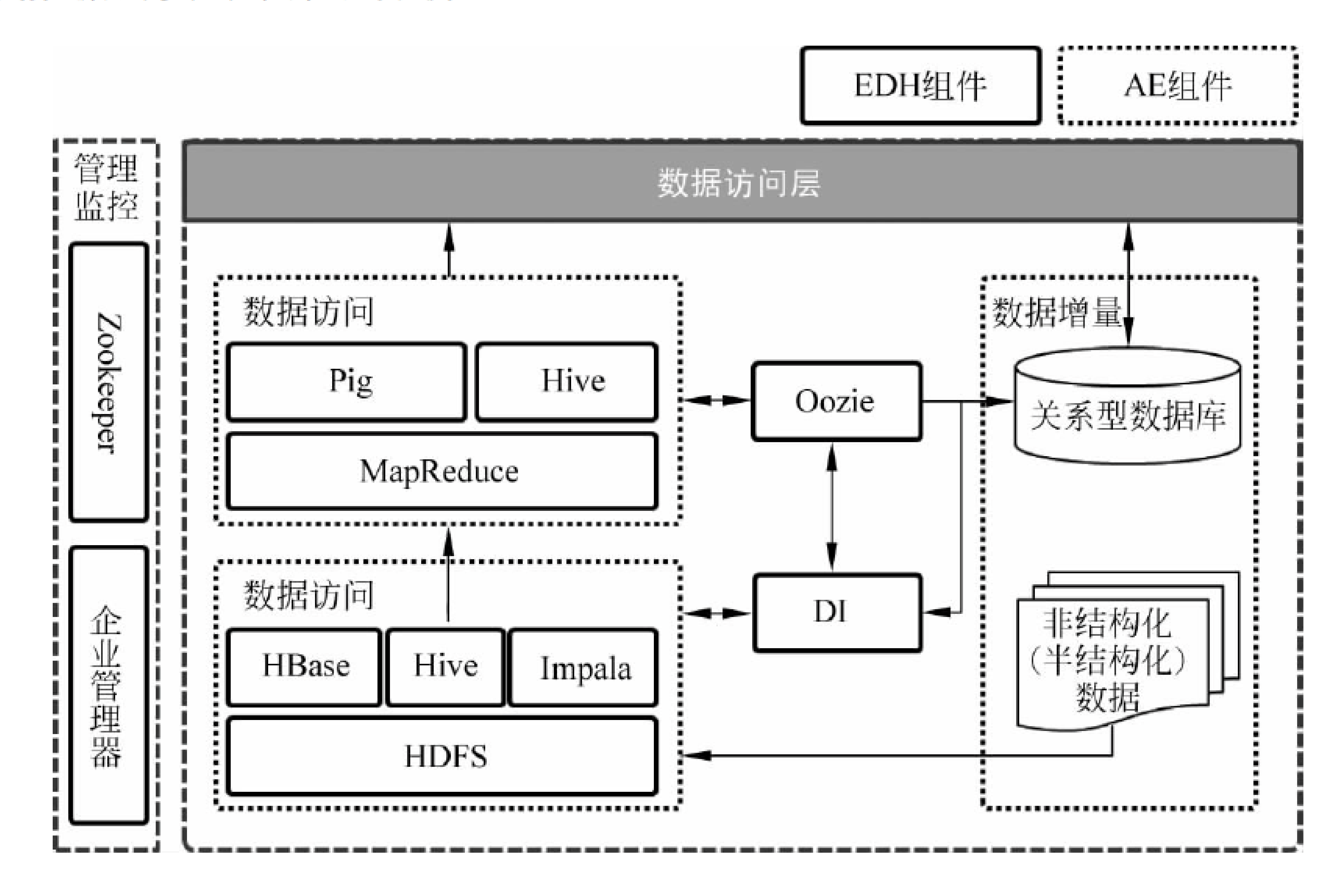


图 7-37 批量处理架构图

3. 大数据集群管理器

企业在利用 Hadoop 处理大数据问题时，面临如下问题。

(1) 部署：前期咨询和需求分析服务欠缺，对 Hadoop 架构普遍陌生。

(2) 应用：缺乏 MapReduce 设计能力，缺少能够提供完成解决方案的专业厂商。

(3) 运维：缺乏有经验的本地支持厂商，系统管理和调优的门槛较高。

上述这些问题制约了 Hadoop 相关技术在企业中的应用和推广，为降低 Hadoop 技术的门槛，快速低成本地应用 Hadoop，EDH 集群系列产品在安装部署、数据处理服务和运维

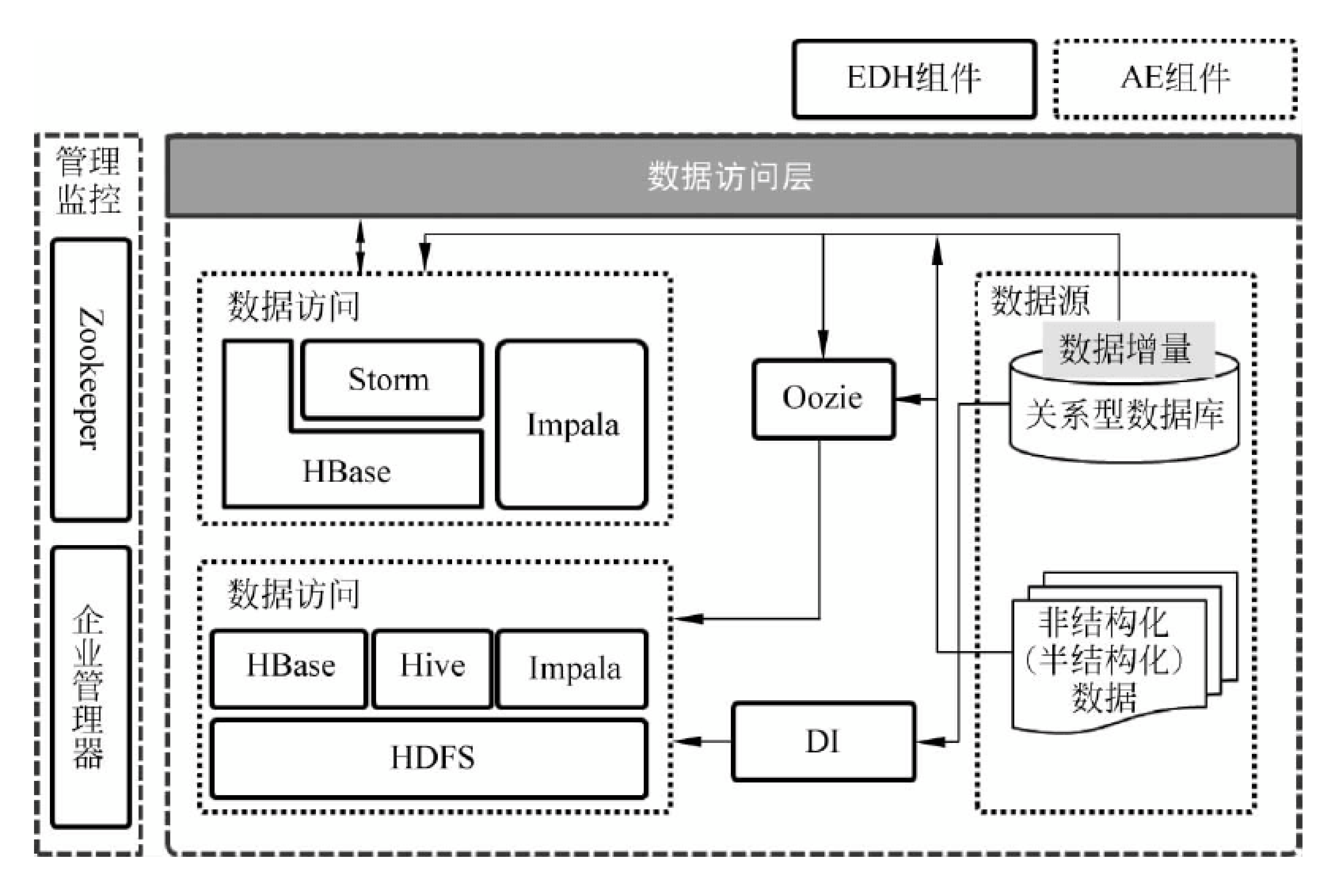


图 7-38 实时处理架构

监控上提供了整套的技术解决方案，其中企业管理器作为实施运营工具，具有重要作用。

EDH 集群管理器主要功能如下。

(1) Hadoop 服务组件和集群主机的监控功能。实现对 HDFS、MapReduce、Hive、Pig、HBase、Zookeeper 等组件的部署、管理和监控以及系统主机 CPU、内存、网络情况的监控，服务组件信息。

主要特征包括：可以根据预置的指标，实时监控服务组件的健康状态；支持可视化的分析展现，能够更好地查看依赖关系和性能指标；通过 RESTful API 对外部系统提供了集成接口；提供失败恢复机制，能够使得各个组件恢复到一致状态；提供授权机制，支持按权限管理用户；支持组件之间消息传输加密；提供详尽的错误堆栈信息，支持日志分析。

(2) 在 EDH 的底层数据处理服务组件的基础上，提供基于 Web 的数据分析处理工具。能够实现：支持数据仓库 Hive 和分析引擎 Impala 的查询编辑、运行、历史保存等；支持 EDH 存储文件的浏览、编辑、下载等；支持数据处理脚本的设计、提交、运行日志等；支持数据处理流程的可视化设置，调度运行等；支持查询数据的简单可视化分析。

7.7.3 主流大数据分析平台

1. Palantir 旗下大数据分析平台

Palantir，提起这家公司就会让人觉得如雷贯耳，曾用大数据帮助 CIA 打败本·拉登的 Palantir Technologies，在 2014 年美国大数据公司收入排行榜中排名第一。Palantir 也被称为大数据行业的印钞机，它的客户包括美国国家安全局(NSA)、美国联邦调查局(FBI)、美国中央情报局(CIA)和很多其他的美国反恐和军事机构。而作为 Palantir 旗下的大数据分析平台 Palantir Gotham 自然也受到了广泛的关注和应用。

Palantir Gotham 将多源数据作为起点，包括很多结构化数据，如日志文件、财务数据表和电子表格等，以及很多非结构化数据，如电子邮件、文件、图片和视频，融合数据为人本模型，通过摒除数据类型和数据容量的限制，将多个相关的源数据整合并绘制为简介、一致的

模型。这个模型一经建立后，数据流就会持续不断地流入 Palantir Gotham 平台。与此同时，相应的安全守则也已建立，只有被授权用户才可登入。这些数据的任何更新都会同步到平台，并且用户进行分析时，他们所有的行为都会被自动记录、归因分析和储存。用户可以通过建立在此平台上的各种综合性应用与数据进行互动。他们可以即刻搜索所有数据源，将数据关系可视化，探索不同的假设，发现未知的关系，揭示隐藏的模式，与同事分享自己的见解。

Palantir Gotham 平台后端集成了一系列功能，它们主要用来整合不同的数据源以进行安全、协同的分析。此时平台扮演着企业知识库的角色，收容着企业全部分析活动的所有记录。Palantir Gotham 平台具有的优点和特性主要如下。

(1) 建模灵活性。Palantir Gotham 平台的数据模型，能快速定义和重定义数据，这让它被称为“动态本体”，同时也让整合不同来源的不同数据为一个整体成为可能，这个过程正符合人们对信息的自然设想。

(2) 隐私和安全控制。平台一开始就设计了隐私保护功能，用来支持精确的数据处理，多层次的安全保护，完全性的审核。用户被分配给不同层级的准入许可，以此来管制他们与数据互动的权利。

(3) 合作。Palantir Gotham 平台支持多样性的合作，包括能够突破跨境机构、功能、地域间限制的合作，连接安全模型和数据模型间的合作，连接低频、高延时下的不同网络、甚至卫星的合作；同时数据的安全性和完整性都有可靠的保障。

(4) 可扩展性，可定制性，应用程序接口。Palantir Gotham 平台每一层的堆栈都被设计成一个完全开放的平台。经由动态本体技术(Dynamic Ontology)整合的数据可以通过 Java 入口作为 Palantir 对象接入。

(5) 知识管理用户可以探索不同方向的推理想法，一路记录下每一步，并可以跳回他们探索过程中的早期节点。同时，数据分析者还可以在不丢失自己工作进度的情况下与他人分享自己的见解。这些便利条件会促成一个版本控制知识库的诞生。它将机构内不同分析者对数据的见解累积起来，并将其转换为数据。在未来，企业可以利用这些分析成果取得杠杆式飞跃。

(6) 算法处理。Palantir Phoenix 工具提供了编译和分析大规模数据集的功能，同时还提供了一个强大而灵活的框架用来实现该功能的自动化。非技术出身的分析师也可以利用种子框架在不用写一行代码的情况下创作出一份精彩的成果。

(7) Palantir Gotham 平台前端提供了一整套的集成工具，这套工具在语义分析、时间分析、地理空间分析、全文分析方面均做了优化。用户可以将数据对象在不同应用之间拖放以获得流畅、全面的分析经验。相关工具和应用程序主要有：图表，地图，对象资源管理器，浏览器，移动端等。

2. IBM Platform Symphony 大数据平台

IBM Platform Symphony 作为可伸缩性极强的企业级网格服务器 SOA 中间件，可用于在可扩展、共享、异构的网格中运行分布式应用服务。它充分利用可用的计算资源，提高并行应用的运行速度并快速得到计算结果，良好地满足数据密集型与计算密集型应用，全面提升系统性能。在全球，IBM Platform Symphony 正在为世界 75％的金融机构提供服务，其中，世界排名前 5 的银行中有三家在使用 IBM Platform Symphony，世界排名前 20 的银行

中有12家在使用IBM Platform Symphony。在中国，中信银行正在应用IBM Platform Symphony满足基于大数据分析的商业与风险管理应用。

作为一个企业级大数据和分析平台，Platform Symphony的一个核心优势是，它能屏蔽底层基础设施的复杂性，在共享底层基础设施环境的基础上，为上层各个不同的大数据应用提供一个多租户的环境。同时，它还能支持Hadoop应用，允许一些基于Hadoop开发的大数据应用和一些并行计算分析应用，在一个集群或者同一个分布式基础设施环境上运行。

以金融领域常见的交叉货币互换期权价值分析应用为例。为了完成这项工作，用户需要模拟未来一段时间内本币利率、外币利率和外汇汇率的发展趋势，并通过用各种不同的利率组合来计算合约在不同情况下的价值。实践中广泛采用蒙特卡罗路径模拟的方式，采用这种分析方法需要模拟大量的蒙特卡罗路径（模拟的路径越多，其精确度越高），计算量非常大，而且耗时。如何管理集群资源，让其并发地完成多个蒙特卡罗路径的模拟，是一个严峻挑战。通过IBM Platform Symphony构建一个分布式网格计算平台，可以帮助客户快速部署、管理、监控资源，并保证计算的并行化，且没有单点故障以提高可靠性，最终快速获得所需要的结果。

Platform Symphony为大数据分析不仅提供了强大的管理、调度和监控功能，同时还提供了很强的对开源软件的支持和兼容能力，不仅让基于Hadoop、Spark开发的应用可以在Platform Symphony中运行，同时能让用户可以用熟悉的开源工具，如IPython、Zeppelin等，来对运行结果进行分析和展现，极大地方便了数据的处理工作，最大化地提供了处理效率。另外值得一提的是，与这些Spark、Hadoop等开源软件相比，由于Platform Symphony是采用商业化的软件模式开发的，因而在性能、时延等诸多方面都比开源产品有明显优势。这也反映在一些实际应用性能测试上，相较开源软件，采用Platform Symphony可以有一些大幅度的提高（有些可能达到数十倍），尤其是一些对时间延迟比较敏感的应用。

3. 清数NEO大数据平台

清数NEO大数据平台是一款面向企业的大数据商业智能产品，提供大数据全链条技术及业务支撑，包括数据清洗处理、数据仓库搭建、数据分析挖掘到最终的数据可视化展示，产品处于国内领先水平。

通过NEO平台，能够帮助企业快速实现大数据运行服务搭建，提供分布式数据库、分布式数据挖掘平台、流计算引擎及相关的自动化运维工具。内嵌的商业智能分析模块IDView，则通过全新的方式，解决企业数据分析难，技术人员任务压力大的问题。无论是销售数据、ERP数据、税务数据还是社交媒体数据、网站访问数据等，都可在IDView中通过单击、拖曳的方式实现数据分析，无须技术人员介入，满足企业快速分析、灵活报表的需求。而针对企业的特定需求，企业也能通过自定义模板、定制化开发等方式，快速实现业务需求，从而推动企业实现数据智能化管理，增强核心竞争力，激活数据，智创未来。

目前，该产品广泛应用于企业私有数据中心建设中，帮助企业打破数据孤岛，实现多种数据分析业务，包括精准营销、销售分析、客户分析、市场监测和预测分析、财务分析、生产及供应链分析、风险分析、质量分析、业务流程等。

在清数NEO大数据产品背后，其运营团队还为多个行业服务提供分析挖掘模板，包括医疗、教育、税务、政务、金融等。企业只需下载模板，即可快速实现相关行业的分析。

（1）清数NEO大数据平台IDManager详解。

① 软硬一体化解决方案，大数据开箱即用。清数 NEO 平台提供一体机解决方案，通过平台预装及硬件优化，用户只需购买一体机，即可实现大数据开箱即用，最大化减少用户信息化建设时间成本，专注于实际业务的快速开展。

② 纯大数据架构，支持无限扩展，如图 7-39 所示。与传统智能分析平台不同，清数 NEO 平台充分考虑大数据问题，采用分布式计算、内存分析、流式计算等多种方式，实现高可扩展架构，当承载的数据增大后，平台可以通过增加新的节点来获得整体性能的提高，从而解决传统架构中数据爆发后，难以承载的问题。

NEO
INCITE DATA
控制面板　节点　服务　命令　数据库　作业管理　系统管理　Thomas Anderson

节点信息　添加节点

#	状态	节点名称	节点IP	角色	计算单元	内存	磁盘	操作
1	✓运行中	test-1	192.168.50.2	["compute","client","storage","master"]	4	3.74 GB	119 GB	删除
2	✓运行中	test-2	192.168.50.8	["compute","client","storage"]	4	3.74 GB	119 GB	删除
3	✓运行中	test-3	192.168.50.10	["compute","client","storage"]	4	3.74 GB	119 GB	删除

图 7-39　NEO 进程状态监控

③ 一目了然的数据监控展示。通过 NEO 平台能够一目了然地看到系统的整体状况、节点状态，以及内存、CPU、硬盘使用率等，包括整个集群的运行情况，如图 7-40 及图 7-41 所示。

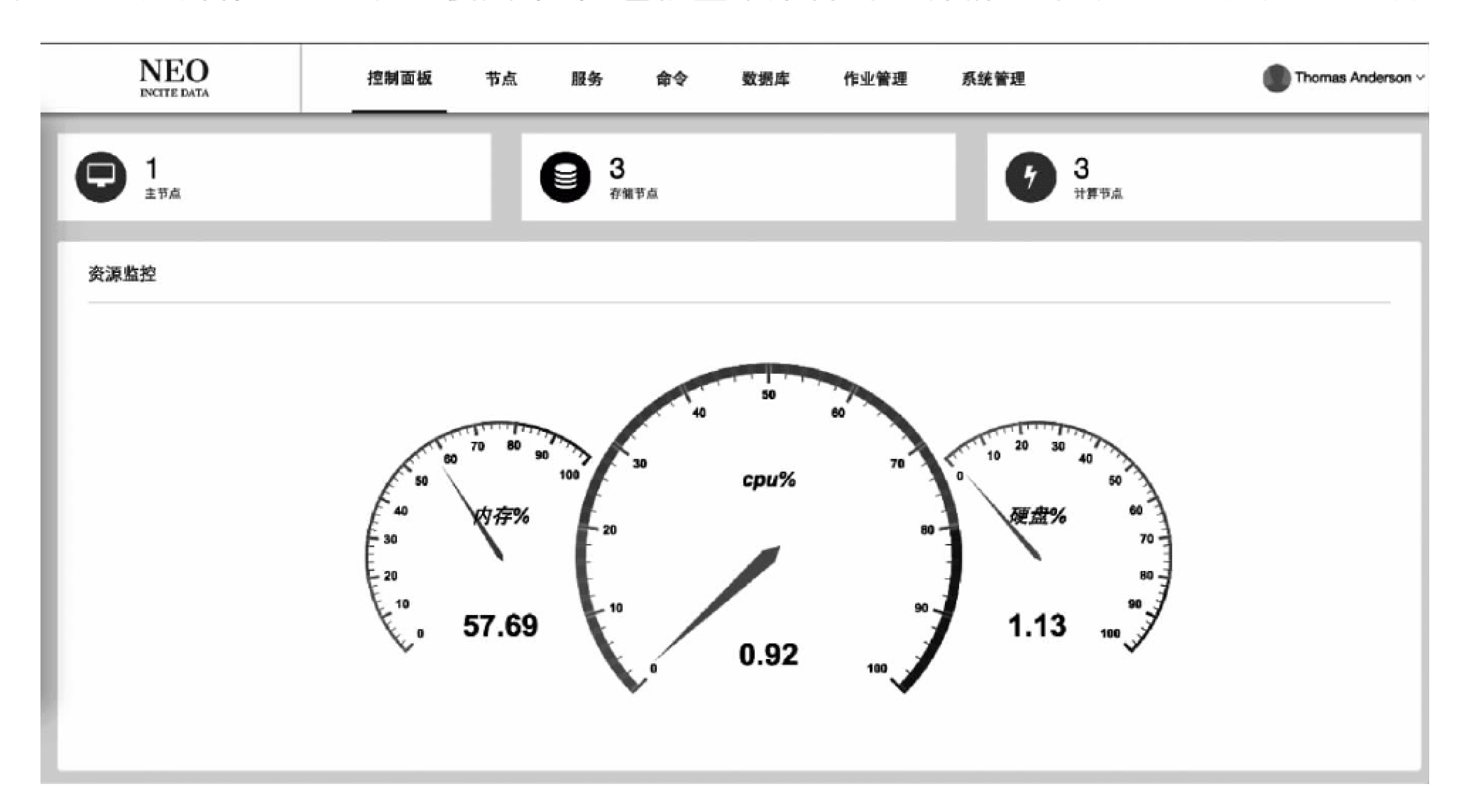


图 7-40　系统运行监控

（2）NEO 大数据平台 IDView 智能 BI 组件详解。

① 原生多行业应用支撑。针对数据分析入门企业，NEO 平台提供多种标准化数据分析模板，通过下载模板，用户能够快速完成行业化标准数据分析，如电商的留存分析、渠道分析、市场数据分析等，帮助用户梳理数据业务，快速实现数据驱动。

② 简便的数据分析、分享方式。清数 NEO 平台旨在为非程序员用户提供快捷的数据分析能力，因此整体使用均采用简单、一目了然的方式。用户只需要通过单击、拖曳相关数

NEO INCITE DATA　控制面板　节点　服务　命令　数据库　作业管理　系统管理　Thomas Anderson

服务信息

#	服务名称	状态	操作
1	分布式锁服务	运行中	▶启动 ■停止
2	分布式文件系统	运行中	▶启动 ■停止
3	资源调度框架	运行中	▶启动 ■停止
4	分布式数据库	已停止	▶启动 ■停止
5	内存计算框架	运行中	▶启动 ■停止
6	监控服务	运行中	▶启动 ■停止
7	ETL服务	!警告	▶启动 ■停止

图 7-41　服务进程控制面板

据字段即可实现多维数据分析、时序分析等，产生的数据分析结果可以通过多种方式进行展示，同时分享给相关人员进行查看。

③ 丰富的可视化展示支持。平台提供 20 种以上的图表数据展示方式，如图 7-42 所示，包括柱形图、条形图、面积图、漏斗图、字符云、标签卡等，同时支持自定义多色彩标记，使得数据结果展示更加直观，帮助企业管理者把握全局，洞察企业问题并发现商机，如图 7-43 和图 7-44 所示，展示行业应用大数据图例。

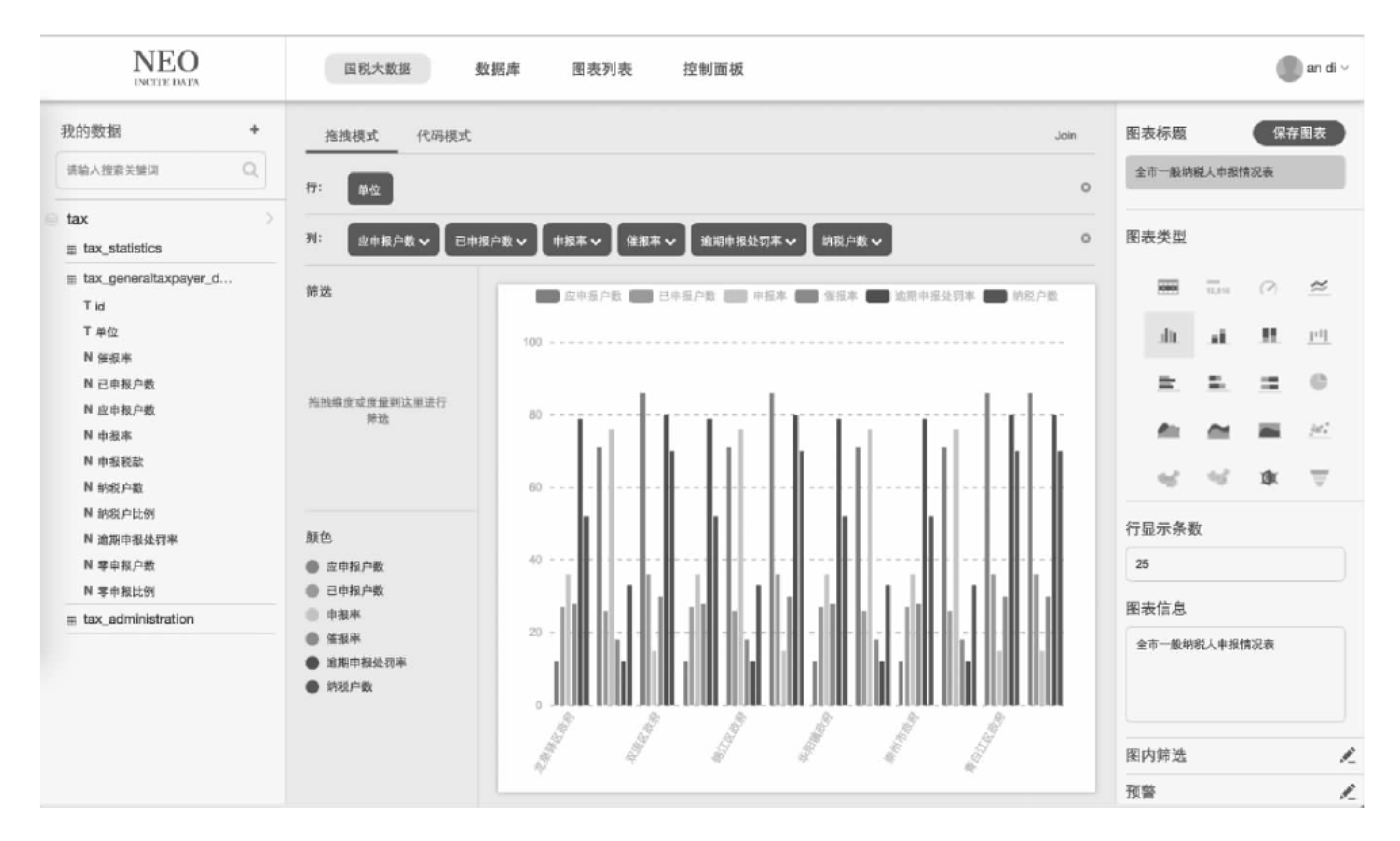


图 7-42　大数据柱状图展示

④ 快速预警条件创建及实时预警分析。以往，在企业数据预警分析中，如库存预警、收入支出预警等场景，往往需要技术部门参与定制开发才能实现，同时，随着告警内容增多，数据复杂度增加，预警往往会出现极大的延时，无法满足业务需求。在清数 NEO 平台中，预警功能的设置则可以通过业务人员完成，基于页面的快速预警条件创建方式，能够让非技术



图 7-43　行业应用大数据展示

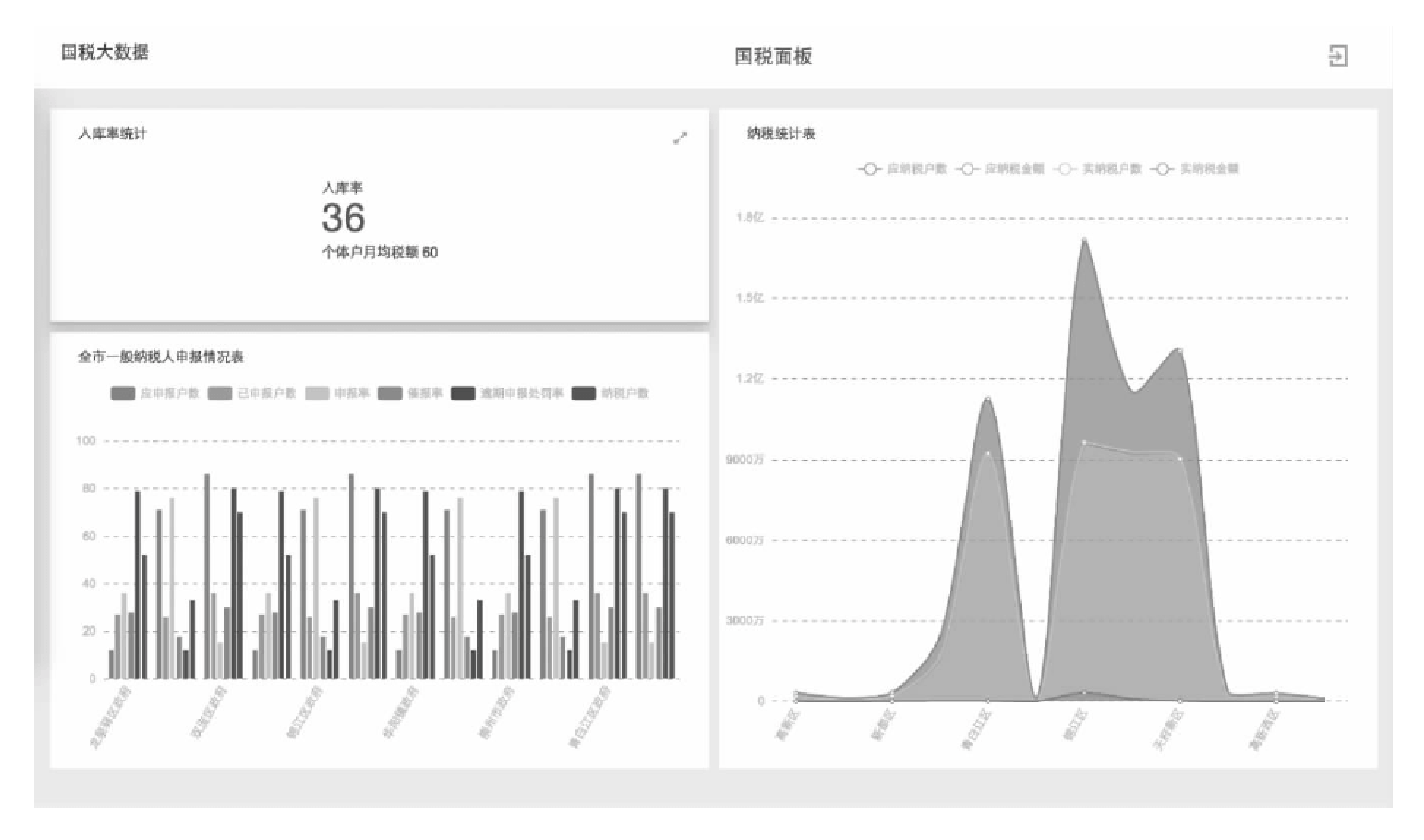
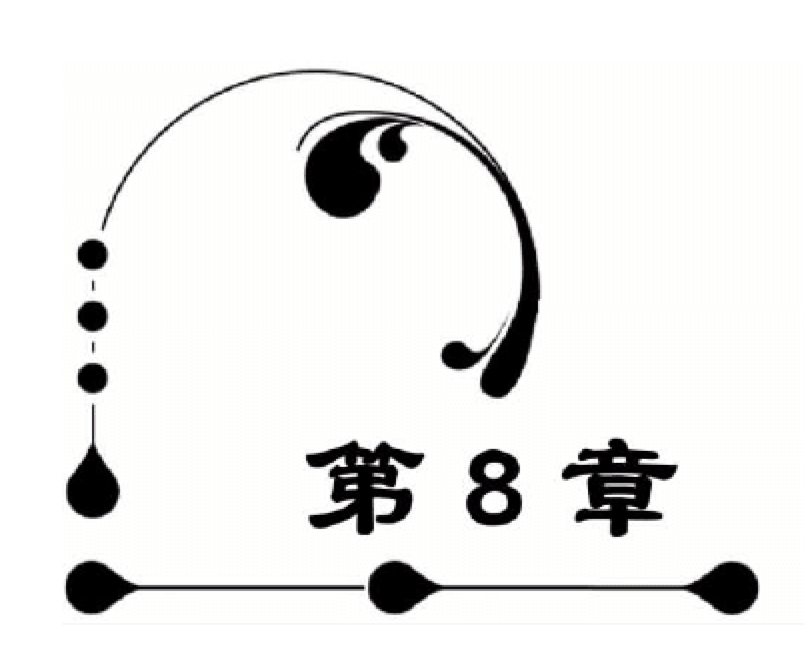


图 7-44　国税大数据展示图例

人员在 10 分钟以内完成任意条件、任意数据的预警设置。另外得益于底层大数据平台的支撑，预警分析通过实时流的方式进行，确保了预警条件触发后，最短时间对外预警。

⑤ 自定义行业面板。摆脱以往固定报表模式，IDView 能够让用户自定义排版数据展示形式，汇聚所有关心的数据，在统一的页面中得到展示。通过分享，相关人员即可看到报表结论。

⑥ 多数据源支持。区别于传统数据仓库架构，NEO 平台充分考虑了移动数据及公开数据的重要性，在传统的数据库支持之上，平台还对多种数据平台提供支持，实现企业内外部数据的整合，提供企业全方位的数据分析。



第8章 医疗健康大数据解决方案

卫生信息化是指以健康信息为核心、管理信息为纽带、分析决策系统信息为主导的全面信息化进程。它体现了现代信息技术在医疗卫生领域的充分应用，有助于实现资源整合、流程优化，降低运行成本，提高服务质量、工作效率和管理水平。众所周知，在都市中奋斗的白领阶级虽然拿着较高的工资，却也付出了极大的心力。据相关统计显示，白领阶层中工作时间超过8小时的高达90%，10小时以上的占62.3%，超过12小时的占20%，而中国白领平均每周的运动时间却只有2.61个小时。长时间超负荷的工作，一再被压缩的运动时间，导致越来越多的白领脱离了健康的“轨道”。由于受限于现有的网络和硬件设施，各区县现有的社区卫生服务应用系统的建设差别较大。比较起来，城区的社区卫生应用软件建设起步早，而在偏远地区，社区卫生服务工作基本停留在手工操作阶段。但是，即使在经济比较发达的城区，各区甚至各社区服务中心都没有统一功能、统一版本的社区卫生服务信息系统，社区服务中心自行开发的应用软件只能满足基本的社区卫生服务要求。这为社区卫生相关政策的执行，社区卫生服务系统与外系统的接口带来了极大的不便。因此，从社区卫生管理的需要出发，急需建设一套保留个性化要求的、全市统一的社区卫生服务信息系统应用软件。

“大数据+医疗”：智慧医疗探索数据显示，当前国内现有2000多款移动医疗APP，且处于快速增长阶段。2014年，我国移动医疗市场规模达到30.1亿元，比2013年增长26.8%，预计2017年将达到125.3亿元。移动医疗APP德国调研公司Research2guidance报告称，当前全球移动健康应用的数量超过10万项，大部分应用的下载量不超过5万次，营收低于1万美元。

《“健康中国2030”规划纲要》提出，要鼓励和规范有关企事业单位开展医疗健康大数据创新应用研究，构建综合健康服务应用。事实上，好多与医疗相关的企业已经开始了这样的探索。以大数据为基础的精准营销，已经在颠覆传统的广告模式。有人说，2013年是大数据元年，未来5年会有一大批基于大数据商业模式的公司催生出来。资深互联网评论人士谢文认为，大数据时代将首先给健康和医疗领域带来深刻变革，因为该领域已经过了思想革命的概念阶段，逐步迈入商业模式创新时期。这或许恰好解释了为何移动健康行业在2016年成为风险投资的热土。如果把大数据时代分为前台、中台和后台三个主战场，前台就是数据终端，负责数据获取和传输，如手机、计算机、智能眼镜、汽车以及各种传感器等，将物质世界和人类社会的一切数据化。在谢文看来，前台是目前争夺的主要战场，出现的创新数不胜数——这正是近两年智能手表、智能手环、电子秤等智能可穿戴设备大热的背景。与此同时，各种健康数据收集平台也在2015年陆续登台亮相：先是三星公司5月底发布一款健康

追踪腕带 Simband 和智能健康追踪平台 SIMI，接着苹果公司在6月 WWDC 大会上发布移动应用平台 HealthKit，数天之后，谷歌紧追不舍在其年度开发者大会上推出名为 Google Fit 的健康平台。近日，微信以公众号为接口，与咕咚、华为、乐心和 iHealth 4 款运动手环展开合作的消息又博到不少中国媒体的眼球。外界纷纷揣测，腾讯公司此举实乃有意借微信打造出一个开放的健康数据平台。面对如火如荼的大数据争夺战，百度董事长兼 CEO 李彦宏 2015 年在黄山召开的"百度联盟峰会"上语惊四座："我们真正想要的数据现在没有，或是还没有搜集上来，已经被搜集上来的数据基本没有价值。"

"大数据＋智能穿戴"：移动医疗创新"戴个手环、弄个眼镜"，计算每天走多少步、消耗了多少卡路里、心跳多少次，对治病没有什么帮助。"互联网公司通过可穿戴设备搜集了很多数据，结果又发现没法对这些数据进行分析"李彦宏说。在利用体检数据方面，美国硅谷早有成功案例。几年前，经尔纬数据技术有限公司创始人糜万军在美国硅谷完成了一个大数据创业项目。该项目利用数据挖掘技术，综合分析斯坦福大学全校员工的体检记录和就诊记录，并据此对所有人每年的医疗费用进行预测。糜万军说，项目成立的初衷，是希望利用个人的医疗信息预测其医疗费用，给保险公司作参考。但后来，美国许多大企业却成为客户的主要来源。变化是这样发生的：糜万军带领的团队，在了解每名员工的健康状况之后，通过数据分析，为其制订了个性化的健身计划，有效地帮助员工改善了健康状况。这项业务受到美国企业的欢迎，从斯坦福大学到思科、苹果等大公司，都乐于购买它的服务。创新总在以极快的速度迭代，但在李彦宏看来，真正能给医疗健康行业带来革新的，是一种"慢数据"：通过一种简单的方法，在三个月、半年甚至更长的时间内，持续不断地监测你的某些指标，通过长时间的数据积累，准确预测你未来患上某种疾病的可能性，以达到中医所讲的"治未病"的效果。这并非空穴来风。2015 年发表在阿尔茨海默症国际会议上的 4 篇论文进一步支持了如下结论：通过对眼睛和嗅觉的检测，能够预测阿尔茨海默症（俗称老年痴呆症）的发生。无独有偶，同年伊利诺斯大学的研究者透露，他们根据现有数据研究发现，人脸的衰老速度与寿命之间存在着确切的关联。假设该研究顺利进入应用阶段，保险公司只需对准顾客的面部乃至照片扫描一番，即可知晓他的夭寿几何，从而优化该顾客的相关保险配置。

"看病难、看病贵"是当前我国一个严重的社会问题，各级医院承担着大部分为人民群众提供优质价廉的医疗服务的任务，任务十分繁重。县级医院是省、市、县城市三级医疗卫生服务网的基础，又是农村三级医疗卫生服务网的龙头，是与群众关系最为密切的公立医院，在我国医疗服务体系中起着承上启下的重要作用。医改提出以信息化建设作为医院改革的技术支撑。大力加强医疗机构的信息化建设，是推进公立医院改革、提高其管理和服务水平的重要手段。经过多年的信息化建设，我国医院信息化建设已经达到一定水平。但是，我国医院信息化建设的发展很不平衡，总体而言，县级医院的信息化建设落后于其他二三级医院，中西部地区医院的信息化建设落后于东部沿海地区医院。这种失衡的状况不利于中西部地区医院整体管理和服务水平的提升，妨碍着医疗信息跨地域的互联共享，目前的状况亟待加以扭转。

为贯彻落实深化医药卫生体制改革精神，中央财政已安排资金，准备在 2015 年启动《2015 年中西部地区县级医院信息化建设项目方案》，重点支持一批县（市、区）医院和新疆生产建设兵团医院的信息化建设。

重庆市作为全国 5 个试点省市之一，承担 4 个国家试点项目建设，是试点项目最多的省

市，是卫生部对重庆工作的肯定和信任。重庆市卫生信息化基础工作与东部发达地区相比有一定差距，试点项目要在重庆能够做出成绩，做成亮点，摸索出在西部地区卫生信息化建设新的机制和体制，为全国卫生信息化建设提供经验，为"健康重庆"建设做出贡献。

8.1 医疗信息化

信息化是现代医疗的发展趋势。医疗信息化是指先进的网络及技术应用于医院及相关医疗机构，实现医疗和管理信息的数字化采集、存储、数据转化与数据整合，以及各项业务流程数字化运作的医疗信息体系。随着互联网、物联网、云计算等的快速发展，特别是数字化医院建设及可穿戴设备、核磁共振、高能射线等的广泛应用，各医疗平台无时无刻不在产生出涵盖人体各部位的成千上万的海量数据，并呈现指数级大爆发，而传统数据库和信息系统架构已无法及时管理和分析这些数据，于是医疗"大数据"时代悄然而至。《大数据时代》作者维克托·迈尔·舍恩伯格，在书中前瞻性地指出，大数据带来的信息风暴正在变革人们的生活、工作和思维，即将开启人们思维、商业和管理重大变革的时代转型。利用这些海量信息资源更好地为临床医疗、医学科研、卫生管理服务，成为当下各级卫生管理机构和管理者亟待决策的时代发展课题，对优化卫生资源配置，促进医疗方式改革，提高医疗服务效率，降低医疗保障成本等具有重要意义。

8.1.1 美国医疗信息化发展情况

全球医疗信息化的开端可以追溯到 20 世纪 50 年代，那时计算机技术也刚刚兴起。然而，相比其他行业的信息化速度，医疗行业的信息化速度要慢很多。这主要与医疗行业所具有的一些特殊性有关。一直以来，医疗技术的发展始终走着一条相对保守的路线，因为医学是一门要求非常严谨的学科，稍微的偏差都可能以患者付出生命为代价。因此，医学的发展更需要精准化、精细化、科学化和现代化。

1. 美国医疗信息化发展回顾

西方最早的"病历记录"可以追溯到公元前 1600 年记录在莎草纸上古埃及的一个手术记录。中世纪著名的伊斯兰医生阿尔·哈兹在中世纪(公元 8 世纪到 9 世纪)延续古希腊科学家记录病情的方式，记录着患者的病情，成为当时记录病例最多的医生。中世纪前的病历记录以医生主观的病情发展和描述为主，主要用于传授医学知识。延从西方欧洲医生保留为自己患者治疗记录的习惯，美国医生在 17 世纪也开始记录和保存为患者治疗的记录，其中的代表就是在爱丁堡接受医学教育的本杰明·拉什医生(Benjamin Rush)。值得一提的是，拉什医生是美国国父之一，他是大陆军的总军医，并且在独立宣言上签过字。

美国第一家开始病历记录的医院是纽约市医院(New York Hospital)。纽约市医院从 1793 年开始对医院的患者进行病历记录，但是病历上只有简单的入院和出院记录和描述。从 1808 年开始，纽约市医院开始把医生个人对于患者病情的记录报告复制抄录下来，作为医院图书馆的档案保存。这个时候的记录就已经包括：病史，病因，治疗方法，以及治疗效果。然而，这个时候的病历不是为患者使用，而是属于医院和医生的私人记录。

从 1821 年开始，麻省总医院的医生也开始记录入院患者的情况，并把这些记录抄录在医院的档案中。在 19 世纪后期，当患者的记录充实并详细后，很多案例才被用于哈佛大学

医学院的教学中。19 世纪后期的病历记录开始包括详细的患者家庭病史、患者生活习惯、病史、身体检查结果、血液和尿液检验结果、病程记录，以及出院诊断和出院说明。但此时的病历仍然内容分散，无序，而且很少有医生的签名确认。例如，研究发现作为 19 世纪名医代表的约翰霍普金斯四大名医之一的“威廉·奥斯勒”(William Osler)的病历记录就非常零散，而且鲜有其签字确认。

现代病历发展史的第一个飞跃式创新发生在 1907 年圣玛丽医院和梅奥诊所。现代医院病历的创始人梅奥诊所的亨瑞·布朗门(Henry Plummer)医生开创了现代病历的改革。现在看来简单易行的纸质病历制度在 20 世纪初期却是医学史上质的飞跃之一。在布朗门医生发明现代医院病历前，病历系统是“流水账”的形式。每个患者没有统一的病历号，而且患者的病程情况和各种检查结果可能分散在不同诊所或医院内。布朗门医生从 1907 年 7 月 1 日开始在梅奥诊所推行新的病历制度，每个患者只有一个病历号和一个集中的病历“夹子”。所有这个患者的病情的记录和各种化验结果都集中在这个“夹子”里，并跟随患者。患者把在不同的医疗机构中就诊的记录放在这个“夹子”里。无论去多少个地方多少次就诊，这个“夹子”包含患者整体的病情状况。很快，布朗门医生发明的这套体系在世界范围内开始普及。此后，1916 年，纽约长老会医院进一步发展并设计了针对每个病区的病历系统。

美国现代病历发展历史的第二次飞跃式创新是对于病历结构标准化。1918 年，“美国外科学会”(American College of Surgery)要求医院对于所有患者情况进行记录，包括对于治疗和结果的总结。

拉里·维德(Larry Weed)在 20 世纪 60 年代带来了美国病历史的第三次飞跃。维德医生将“问题导向型病历记录”(Problem Oriented Medical Record)引入医疗实践中，并将纸质病历电子化。维德医生被誉为“问题导向型病历之父”，其创新在于“问题导向型病历记录”可以让第三方独立地确认诊断。医疗信息化科研和教育非盈利性机构 Regenstreif Institute 在 1972 年推出了第一个电子病历系统。

2. 美国大数据医疗服务模式发展现状

美国政府将大数据定义为“未来的新财富，价值堪比石油”，将“大数据战略”上升为国家意志，投入巨资拉动大数据相关产业发展。2012 年，奥巴马政府宣布“大数据研究和发展计划”，研发大数据技术。

(1) 实施精准医疗计划。

精准医疗是一种基于患者“定制”的医疗模式，在这种模式下，医疗的决策、实施等都针对每一个患者个体特征而制定，疾病的诊断和治疗在合理选择患者自己的遗传、分子或细胞学信息的基础上进行，因人因病而异，是“个体化医疗”的延伸。从概念上可以看出，患者个人的遗传信息(基因组)是精准医疗的支撑基础，也就是对基因组信息的详细注释，以及临床化使用，才能保证精准医疗的实施。精准医疗所使用的工具，通常包括分子诊断、影像以及相应的软件等。2015 年，奥巴马在国情咨文演讲中宣布的精准医疗计划，是新的大规模研发项目，白宫官网发布精准医疗计划的相关细节：2016 年，美国财政预算计划拨付给 NIH、美国食品药品监督管理局(FDA)、美国国家医疗信息技术协调办公室(ONC)等机构共 2.15 亿美元用于资助这方面的科学研究、创新发展。毫无疑问，该投资计划将加快在基因组层面对疾病的认识，并将最新最好的技术、知识和治疗方法提供给临床医师，使医师能够准确了解病因，有针对性地选择用药，避免浪费，减少相应副作用的产生。据说个人的基因筛查成

本已经降到70美元，且成本还在快速下降，筛查速度还在提高，并已有部分能用于临床。由于基因筛查技术的进展，未来的精准医疗有望变成临床现实。

（2）计算机医师临床诊断。

20世纪70年代，美国匹兹堡大学的研究人员开发了用于诊断普内科复杂病症的软件“快捷医疗参考”，这款医疗诊断专家系统能够诊断超过600种疾病，收集了4300种临床表征（包括病情症状、医师问诊、实验室检验结果等），经过系统程序运算，提高快速诊断的可能性。20世纪80年代，美国麻省总医院（MGH）开发和完善了DxPlan项目，其所涵盖的知识领域包括内科各专科的多数疾病及临床表征，使用者可向计算机咨询下一步应做何种检验及测试，以最少的花费得到最多的信息。2010年秋，“伊莎贝尔保健系统”在美国佛罗里达州的奥兰多保健医院联网使用，为医师提供可靠的诊断和治疗建议，一些经验较少、临床实践不多的医师能从该系统获得更多帮助。2013年，由IBM的30位工程师耗时3年研发的计算机医师沃森（Watson）在美国安德森癌症中心开始上岗，其既是癌症诊断专家，又是医疗服务管理的专业人士，从此计算机辅助诊断翻开新的一页，正式迈入“沃森时代”。沃森的运作模式非常类似人脑，自然语言处理能力能全方位地模仿人类的医师，能像真人医师一样“当面”听取患者对疾病的叙述，再对患者的问题进行解答，然后做出诊断和开出药方。沃森具有超强的认知计算能力，从患者病例和丰富的研究资料库中寻找资料，为临床医师提供有价值的见解，帮助医务人员找到最有效的治疗方案，在医疗领域具有广泛的应用。

（3）建立患者为中心的医疗模式。

2013年《美国医学会杂志》（*JAMA*）撰文指出，大数据在医疗方面的应用势不可挡，将从新知识的产生、医疗质量的提高、个体化医疗和临床决策等多个层面，推动医疗模式从以医师为中心向以患者为中心的改变。以患者为中心的高效医疗模式代表医疗服务发展和服务理念的转变，是医疗体制改革的最终目标。以患者为中心的医疗模式充分尊重患者，对其兴趣、需求和价值观做出快速回应，确保所有临床决策以患者的价值观为导向。而尊重患者的价值观、个体化特征和需求，协调和整合不同专业的医疗服务、情感支持，做出决策时征求患者和家属的意见，保持医疗服务的连续性和可及性，是提高医疗质量的基本要求。大数据则因为有效的数据整合模式，可以满足以患者为中心医疗服务的个性化医疗、协调和沟通、患者支持和赋权以及良好可及性等多方面需求，为其提供卓越的技术平台，从医学研究、临床决策、疾病管理、患者参与及医疗卫生决策等方面推动医疗模式的转变。

在2016年世界生命科学大会召开之际，中共中央政治局常委、国务院总理李克强做出重要批示。批示指出：生命科学是21世纪重要的综合性学科领域，关系人类的生存、健康和可持续发展。中国政府正在深入实施创新驱动发展战略，落实“健康中国2030”规划纲要，通过科技创新有力推动生命科学领域的研究与相关产业快速发展，对提高人民健康和生活水平、改善环境质量正发挥着日益重要和明显的作用。希望中国科学家、企业家与各国同行一起，围绕本次世界生命科学大会的主题，瞄准生命科学重大需求，进一步加强交流与合作，相互借鉴，以更多科学突破和创新积极应对人类生存发展面临的共同挑战，形成新的生产力，推动世界经济社会可持续发展，共创人类美好的未来。2016世界生命科学大会在北京开幕，10位诺贝尔奖获得者、3位世界粮食奖获得者、3位沃尔夫农业奖获得者齐聚。生命科学大会围绕目前全球的热点领域，如**精准肿瘤学、免疫治疗、基因编辑、干细胞与再生医学**等进行了为期3天的分组讨论。涉及主题多达66个，是迄今为止我国举办的生命科学领

域层次最高、覆盖面最广的一次国际学术盛会。

8.1.2 我国医疗信息化发展趋势

数字化医院是我国现代医疗发展的趋势。“数字化医院”是指将先进的网络及数字技术应用于医院及相关医疗工作，实现医院内部医疗和管理信息的数字化采集、存储、传输及后处理，以及各项业务流程数字化运作的医院信息体系。“数字化医院”是由数字化医疗设备、计算机网络平台和医院业务软件所组成的三位一体的综合信息系统。数字化医院工程体现了现代信息技术在医疗卫生领域的充分应用，有助于医院实现资源整合、流程优化，降低运行成本，提高服务质量、工作效率和管理水平。

1. 我国医疗信息化发展基本情况

数字化医院一般由以下系统组成：HIS(Hospital Information System，医院信息系统)、PACS(Picture Archiving and Communication Systems，医学图像档案管理和通信系统)、EMR (Electronic Medical Record，电子病历系统)、LIS(Laboratory Information System，检验信息系统)、CIS(Clinic Information System，临床管理信息系统)、RIS(Radiology Information System，放射科信息系统)、EHR(Electronic Health Record，电子健康档案系统)、GMIS(Globe Medical Information Service，区域医疗卫生服务)。

医院管理信息系统(HIS)是以财务为中心的，偏重管理。“医院信息系统”是医院的管理中枢，包含财务、人事、住院、药品、门诊、医技、病程、收费等多个子系统，同时承担着“临床管理”与“行政管理”的双重使命。

临床管理信息系统(CIS)是偏重于病人信息的，更加倾向于医疗相关的信息。人们常常把关于病人化验信息、放射的信息和病人临床检查信息，划归临床信息系统。

医疗影像系统(PACS)是医院的影像中心，它承担着从CT、X光机等各类成像检查设备中采集影像资料、对这些资料加以处理和存储、并为一线医师提供查询服务的使命。

电子病历系统(EMR)是医院的病历中心，它详细记录了患者的治疗方案和治疗过程，既为医院积累了宝贵的治疗经验，又为处理医患纠纷提供了不可或缺的证明文件。

社会保险系统则连通了医院与社保部门的业务后台，它为医院接诊并服务好广大社保患者提供了支持。

数字化将推动医院集团化、区域化，并改变医院原有的工作模式。建立区域性的影像中心(病理、CT、MRI等)实现医学图像网络传输。建立区域性的中心实验室实现检查结果网上传输，节约资源。信息中心社会化，医院不再建立网络、服务器中心，将采用租用电信运营商网络线路，建立区域性的数据中心、服务器中心和数据仓库。实现医学文献资料的共享，解决各医院网络建设重复、利用率低、资源浪费的缺陷。区域性的各类医学服务中心的建立，将使卫生资源获得最大程度的利用。

信息系统建设作为医疗行业信息化的核心内容，在近几年的发展中经历了不同的阶段。目前，中国大部分的医院信息系统仍然是以经济核算为中心的管理信息系统(HIS)，仅有小部分的医院在管理信息系统的基础上开始建立用于临床医疗业务的临床信息系统(CIS)，并且系统建设主要还是集中在大中型医院。临床信息系统的主要功能是支持医院医护人员的临床活动，收集和处理病人的临床医疗信息，丰富和积累临床医学知识，并提供临床咨询、辅助诊疗、辅助临床决策，提高医护人员的工作效率。广义上的临床信息系统包括医生工作

站系统、护理信息系统、检验信息系统(LIS)、放射信息系统(RIS)、医疗影像存储与传输系统(PACS),以及电子病历(EMR)系统等。

2. 我国大数据医疗服务模式快速发展

中国在互联网技术、产业、应用以及跨界融合等方面取得快速进展,打开了医疗卫生体制变革的无限可能性和想象空间。国务院 2015 年 7 月颁发的《关于积极推进“互联网+”行动的指导意见》清晰勾勒了“互联网+医疗”的行动路线,对处于火热、混沌中的医疗改革不仅是催化剂,更是一股提神醒脑的清风。

(1) 在线医疗。

“互联网+”是把互联网的创新成果与经济社会各领域深度融合,推动技术进步、效率提升和组织变革,提升实体经济创新力和生产力,形成更广泛的以互联网为基础设施和创新要素的经济社会发展新形态。史上最具互联网思维的《指导意见》,确立了“到 2018 年在健康医疗领域互联网应用更加丰富,公共服务更加多元,社会服务资源配置不断优化”的发展目标,并明确指出:发展基于互联网的医疗卫生服务,支持第三方机构构建医学影像、健康档案、检验报告、电子病历等医疗信息共享服务平台,逐步建立跨医院的医疗数据共享交换标准体系。各医疗机构要积极利用移动互联网提供在线预约诊疗、候诊提醒、划价缴费、诊疗报告查询、药品配送等便捷服务。引导医疗机构面向中小城市和农村地区开展基层检查、上级诊断等远程医疗服务。鼓励互联网企业与医疗机构合作建立医疗网络信息平台,加强区域医疗卫生服务资源整合,充分利用互联网、大数据等手段,提高重大疾病和突发公共卫生事件防控能力。积极探索互联网延伸医嘱、电子处方等网络医疗健康服务应用。鼓励有资质的医学检验机构、医疗服务机构联合互联网企业,发展基因检测、疾病预防等健康服务模式。“互联网+医疗”的最大优势是能够实现健康与疾病诊治相关信息采集、储存、交换,以及共享使用全过程的自动化和智能化,提高优质医疗资源的可及性,解决医疗行业缺乏标准和规范、缺乏连续性、容易出现重复诊断和治疗等低效率问题,促进“最佳医疗实践”的推广。我国有完整的公立医疗体系,容易实现信息的互联、互通和互享,创造出远程医疗、移动医疗、可穿戴设备等更多为患者服务的新模式。

(2) 移动互联网医院群。

2015 年 7 月深圳市南山区卫生计生局及区属 5 家医院作为“互联网+”的主体,成功地向公众开放“移动互联网医院群”暨“南山看病易”服务平台,整合区域医疗资源,构建从社区到医院、从门诊到住院、从医疗到健康的全流程便民惠民服务体系。打破医院与患者的物理围墙,打破各医院间的信息壁垒,将医院及社区健康中心自身的服务延伸到移动互联网,无须下载与安装 APP,只需在微信公众号关注“南山看病易”,就能获取初诊、智能导诊、预约挂号、门诊付费、检验/检查报告、住院押金预交、住院每日费用清单、出院小结、住院结算、就医评价、健康资讯等,降低民众获取医疗与健康服务的门槛,节省患者看病就医时间,改善患者就医体验,为医疗服务的数据互联互通、个人健康档案信息共享、医疗大数据的积累打下坚实基础。移动互联网医院群具有 5 大领先优势:①全国首个“1+N”模式的“移动互联网医院群”平台架构,统一部署“移动互联网医院”平台,同时构建“南山看病易”区域统一入口及 5 家医院独立入口,确保在区内任意一家医院均可享受到一致的就医体验,实现区内医疗资源的互补与共享,方便区内市民选择就医;②一处建卡、全区就诊,患者在医院群的任意一家医院登记建卡,便可在区内的所有医院实现一卡通就诊,无须重复建卡;③区域检验、

检查结果的共享，患者在任意一家医院的检验、检查结果信息，均可在群内医院调阅与共享；④社区初诊引导，为分级诊疗打下基础，将社区健康服务中心的医疗资源接入群内，在导诊环节引导患者到社区健康服务中心进行初诊，为未来完善分级诊疗、双向转诊、区域协同打下坚实的基础；⑤区域私有云部署、安全自主可控，在区域卫生信息中心基于 Apusic 自主知识产权的中间件构建私有云，部署患者移动服务平台，既满足医院群间的信息集中共享，又达到安全自主可控，这是移动互联网医院群平台与其他第三方 APP 的本质区别。其将建设区域智慧医疗体系，形成更好的分级诊疗、双向转诊、区内协作，让民众切实感受到医改的成效。

(3)“云医院”运营模式。

我国患者看病难、看病贵问题长时间得不到有效解决，医患关系紧张，医师工作强度大、收入低、风险高，三甲医院超负荷运作，而基层医疗机构利用效率却不高，互联网则是最可能迅速改变这一状态的切入口。2015 年 3 月，全国第一家云医院——“宁波云医院”正式启动运营，这个基于云计算、大数据、互联网、物联网等新一代信息技术的城市健康平台，正试图在解决现有医疗卫生系统性问题的同时，用互联网手段放大现有医疗资源，成为面向全世界的“无围墙”的医院联合体。据宁波市卫生和计划生育委员会信息，宁波云医院将成为一个集健康大数据采集、健康管理、医疗、康复服务等为一体的协同医疗与健康管理平台，帮助医院提升现有的医疗服务效率，开拓健康医疗服务更大发展空间，通过互联网完成大医院与基层医院、知名专家与社区医师、医师与患者之间的互动与沟通，实现跨区域、资源共享、协同的医疗服务模式。其既是一个医师多点执业的平台，也同时是一个集成的相关产业平台，首批接入“宁波云医院”平台的基层医疗机构共 100 家，签约专科医师、家庭医师 226 名，首期在“宁波云医院”线上开设高血压、糖尿病、心理咨询、全科医师等 4 个“云诊室”。此外，“云医院”已经与本地连锁药店等第三方机构实现互联，“云医生”线上处方可以方便地流转到连锁药店，患者可以根据实际情况就近取药或享受配送服务。可见，这个云医院线上是一家虚拟医院，线下是一家混合所有制医院，线上、线下既能实现门诊、住院、检查、体检的预约服务，又能实现定制的健康管理、咨询、干预与指导，对特定人群、特定病种实现规定范围内的诊疗。云医院平台将与电子健康档案协作平台、区域医疗服务平台协同服务，实现民众电子健康档案共享调阅和检验、检查远程诊断。民众可通过网上支付和网上药店，足不出户就能购买到高品质的医疗服务。

8.1.3 医疗健康大数据挑战和机遇

基于大数据的医疗服务模式创新，有赖于新技术对海量关联性数据的整合分析，发现独立数据系统不可能发现的有价值的信息。由于我国医疗卫生面临资源配置低效的问题，加之互联网医疗起步较晚，医疗卫生数据的挖掘分析面临着诸多的问题与挑战。

1. 医疗数据整合

医疗领域大数据覆盖医院、区域医疗中心、医疗保险公司、药物管理分析单位、医疗设备监控中心等，数据资源分散在不同的数据池中，包括电子病历、结算与费用数据，医疗厂商的医药、医械数据，医学研究的学术数据，区域卫生信息台采集的居民健康档案，政府调查的人口与公共卫生数据等，彼此之间没有太多联系。同时，医疗数据主要产生于搜索引擎、社交网络、通话记录、传感器等，数据格式如文本、日志、图像、视频、机器数据等结构化、半结构

化、非结构化数据多种多样，且随着可穿戴、PET 等先进医疗设备的广泛应用，非结构化数据快速增长，占总量的 70％～80％。将不同来源、不同类型、不同领域的数据进行转换清洗、重组整合，消除“信息孤岛”、打通衔接通道、建立协作共享机制、激活休眠数据的潜在价值，是一个亟待处理的现实问题。

2. 医疗数据存储

大数据更强调数据的完整性，大量而非精确、非结构化数据进入数据样本，传统的数据库难以实现有效的存储和加工。一是容量问题，“大容量”通常可达到 PB 级的数据规模，因此，海量数据存储系统一定要有相应等级的扩展能力；除数据规模巨大之外，还拥有庞大的文件数量，因此，如何管理文件系统层累积的元数据也是一个难题；二是延迟问题，医疗大数据应用存在实时性问题，需对数据进行实时或准实时的处理、秒级的查询需求响应；三是数据库问题，医疗大数据也是非结构化数据，传统的结构化数据库已经无法满足存储要求，需升级医院数据库系统。这些问题的解决必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术，而这些新技术的采用和实现是一个复杂、长期的系统性工程，难以一蹴而就。

3. 医疗数据挖掘

大数据技术的战略意义不仅在于掌握庞大的数据资源，而且要对这些数据做出快速的专业化处理。传统的医疗诊断主要以检验报告为手段，相当于数据的初次或直接利用，是一种“平面化”的分析。数据挖掘的主要方式既包括报告诊断，也包括数据建模和潜在知识的挖掘，相当于数据的二次利用或间接利用，是一种“立体式”的分析。初次利用包括信息调阅共享、卫生服务智能提示与诊断辅助，还有各类基于信息共享的业务协同服务等。二次利用主要是根据卫生行政与管理需求实现的 BI 统计、绩效分析等。医疗数据挖掘就是要根据不同的医疗管理目标和服务需求，使用不同的数据分析技术和工具，不仅要为传统医疗诊断分析搭建更好的信息平台，而且要使医疗诊断分析更加准确、权威、及时和高效。怎样对大量非结构化数据进行有效的数据挖掘也是医疗的难题之一。

4. 医疗数据检索

大数据技术是通过非常快速地采集、发现和分析，从大量多类别的数据中提取价值的新一代 IT 技术与架构。由于大数据的 5V 特性，对大数据进行检索就和在互联网上用百度、Google 进行检索一样复杂，传统的检索方式为关键词检索，但这种检索方式往往检索出大量无关的信息，无法满足大数据的检索要求。语义技术作为一种检索准确性较高的新技术出现在人们面前，微软已经将其用于互联网检索并建立起自己的检索引擎 Bing，将这类新型的检索技术有效地运用于医疗大数据的检索也是一项艰巨的任务。

8.2 医疗健康大数据综述

随着公共卫生领域的信息化建设，各地政府和公共卫生医疗行业都将医疗行业数据处理及共享作为信息化建设的重点之一，而随着国内首轮公共卫生医疗行业信息建设浪潮而来的是庞大的医疗信息，以及医疗系统间的信息不对称、不共享。同时，卫生数据分散在各医疗机构的信息系统中，与中心平台数据要求存在结构不同、标准不统一的情况，传统的由医疗机构系统开发商完成数据上报的方式存在医疗数据处理工作量大、项目开发进度缓慢、

数据质量低、医疗单位协调难、数据监管难等问题。

在这样的医疗行业需求背景下，随着大数据处理思想的引入，上述难题将得到有效解决。通过对公共卫生大数据的获取和分析，并将数据与各级医疗平台进行实时共享，对分散医疗卫生机构的数据进行快速、有效、可靠的采集，实现医疗卫生机构卫生数据的有效接入，已成为医疗大数据项目建设的重点与难点，将对公共卫生医疗信息化建设起到至关重要的作用。

在医疗数据处理工作量方面，医疗大数据处理机制解决协同各业务系统维护人员根据平台要求，提供数据查询脚本，进行数据采集、转换程序的编写与调试，医院端大量的业务分析工作成为制约数据有效上报的瓶颈。在医疗项目开发进度方面，传统的平台开发商需要与多个医疗机构明确接口内容、传输协议、联调测试，导致平台数据共享与交换部分的开发周期长、进度慢，而独立的医疗大数据提供商独立于平台开发商与医疗机构，从独立第三方角度来汇聚来自医疗机构的大数据，并将其进行统一的数据质量处理后上报给平台，从而在保障数据质量的基础上，加快了项目建设周期。另一方面，传统的医疗数据监管需要靠人工方式进行汇报，无法对问题及时发现、及时解决，很难实现对医疗系统数据的整体有效监管，在大数据背景下，其海量数据的监管需依靠自动化、智能化的方式进行统一的集中监管，以便于及时发现、定位、追溯、跟踪和解决问题，从而降低数据监管难度和成本。

8.2.1　医疗健康大数据类型

通常所说的医疗大数据指的就是医院医疗大数据，这是最主要的医疗健康大数据，产生于医院常规临床诊治、科研和管理过程，包括各种门急诊记录、住院记录、影像记录、实验室记录、用药记录、手术记录、随访记录和医保数据等。这些医疗数据中的大多数都是用医学专业方式记录下来的，以临床实践自然随机形式存在，是最原始的临床记录。从临床管理或研究角度看，这些数据是关于病人就医过程的真实记录，或者也可以说是临床医疗行为留存的痕迹，每一个数据都具有价值，包括记录不完善或错误的数据，都可能隐藏了有待发掘和利用的重要医学信息。如图 8-1 所示为医疗健康大数据类型图。



图 8-1　医疗健康大数据类型图

8.2.2 临床服务数据

临床数据主要包括综合电子健康纵向记录，如诊断、问题列表，现在和过去的药物，结果测试以及病人各自所在的治疗单位及其病人所接触的设施等。它们成为临床决策支持系统和大数据分析系统的基础。

8.2.3 公共卫生调查和监测数据

区域协同背景下的大数据是重要的医疗健康大数据之一，也是未来医疗健康大数据的发展方向。一方面，区域协同通过医疗健康服务平台汇集整合了区域内很多家医院和相关医疗机构的医疗健康数据，致使数据量大幅度增加。另一方面，由于平台数据收集事先都经过充分的科学论证和规划，所以会比单独医院的数据更为规范。

8.2.4 医学研究性数据

除了上述原生态医疗大数据以外，另有一些医疗健康大数据来自于专门设计的基于大量人群的医学研究或疾病监测。例如，原卫生部近年开展的脑卒中筛查与防治项目，计划在全国各地筛检 100 万脑卒中高危人群，随后对其疾病及治疗进行长期追踪。另一项近年刚启动的重大专项研究是中国环境与遗传因素及其交互作用对冠心病和缺血性脑卒中影响的超大型队列研究，包括了 50 余万人的自然人群，评估遗传和环境危险因素及其复杂的交互作用。专项设计的大数据还包括各种全国性抽样调查和疾病监测数据，如全国营养和健康调查、出生缺陷监测研究、传染病及肿瘤登记报告等数据。

8.2.5 个人健康数据

基于移动物联网的个人身体体征和活动的自我量化数据是一种新型的医疗健康大数据。自我量化数据所包含的血压、心跳、血糖、呼吸、睡眠、体育锻炼等信息，除了有利于帮助人们及时了解自身健康状况外，经过一定时期累积在医学上会变得很有用，既有助于识别疾病病因或防控疾病，也有助于个性化临床诊疗，从而塑造一种新的医疗或健康管理模式。

大数据的另一个资源是远程病人监护。远程病人监控，不仅可以产生针对个人行为的实时数据，而且可以产生针对行为模式和相关治疗的实时数据。产生的数据需要能够处理大量信息系统，特别是在需要远程传播视觉成像的时候，这使得疾病的监测变得更加容易，而且也提供了分析领域的业务增长机会。

生物信息大数据是一类比较特殊的医疗健康大数据。这类数据有很强的生物专业性，主要是关于生物标本和基因测序的信息。虽然在信息内容表达方式上，生物信息大数据与上述所有大数据大不相同，但它直接来源于人体生物标本，并且关系到临床的个性化诊疗及精准医疗，所以可归于医疗健康大数据一类。基因测序又称 DNA 测序，能够从人体组织、细胞、血液或唾液中测定基因全序列。全基因组测序的意义在于能揭示一个人的生命密码。据估计，人类基因测序一次，产生的数据量就可高达 100～600GB 左右。目前，每年全球产生的生物数据总量已达 EB 级，使得生命科学已经成为大数据科学。

8.3 医疗健康大数据总体架构

8.3.1 建设原则

医疗健康大数据建设遵循以下 5 方面的原则。

1. 集中存储

降低业务系统的复杂度，降低故障风险，降低数据丢失风险，提高管理效率。

2. 分层存储

针对不同业务系统数据的特性，将数据分布在 SSD、SAS 和 SATA 磁盘上，最大化地提高系统运行效率，降低建设成本。

根据数据访问频度，自动调整数据存储位置，最大化地提升系统整体性能，智能化加快医院业务流程，提升医疗 IT 效率。

3. 统一备份

完整的每日数据备份，有助于在灾难发生时，提供最近时间点的数据备份恢复能力，降低数据丢失风险。

数据远程镜像和 CDP（持续数据保护）不能作为备份的替代解决方案。

4. 业务连续性

医院业务系统需要 7×24 小时不间断运行，一旦应用系统服务器发生故障，将导致整个医院业务系统中断。

服务器系统集群能够使业务系统主机在发生故障时，将业务切换到备用主机系统继续提供服务，确保医院业务系统的高可靠性运行。

容灾系统能够使医院信息系统在主运营中心发生灾难时，快速地在容灾中心恢复医院的业务系统，将故障恢复时间降到最短。

5. 虚拟化

虚拟化能极大地降低医院信息中心服务器系统的结构复杂度，降低管理难度，降低运营成本。

在容灾中心建立虚拟化服务系统，有助于快速恢复业务系统，缩短系统恢复时间，降低容灾中心建设成本。

虚拟化系统的 V-Motion（虚拟机自动迁移）功能能够提供业务系统的安全运行级别。

利用虚拟化技术，能够帮助医院建立双活数据中心，确保医院业务系统实现真正的无中断和业务系统连续性。

8.3.2 建设目标

县级医院信息化建设的主要建设目标如下。

（1）对于尚未建立医院信息系统的医院，争取能够建立覆盖全院的、以经济核算为核心的管理信息系统。项目实施后，将使医院实现初步的信息化，能够规范收费，算清账目，医院管理水平得到明显提升。

(2) 对于已经建有管理信息系统的医院,支持其建设向临床应用延伸。项目实施后,将使医院的医疗业务获益,不仅有助于就医流程的优化,提高服务效率,也对减少医疗差错、改善医疗质量有所帮助。

(3) 对于少数信息化建设较好的医院,重点推动电子病历系统的建设、各信息系统的集成以及临床路径的应用。项目的实施将进一步规范医生的诊疗行为,不仅有益于提高医疗质量,更为未来实现医疗信息的区域共享、降低百姓就医成本打下基础。

数据中心的核心功能是承担数据存储的任务,另外还要为大数据的机器学习、数据挖掘提供平台支持。针对当前医院数据中心的实际需求,基于 Hadoop 的医院数据中心系统的设计开发主要有以下目标。

(1) 实现数据安全、可靠的存储。这主要由 Hadoop 本身特性保证,在开发过程中,针对医院数据中心实际需求,对 Hadoop 做出个性化改进,以更好地适应医院数据存储。

(2) 与现有数据中心相比,提供更快的数据存储速度。基于 Hadoop 构建的数据中心采用分布式文件系统,数据读写并行执行,极大地提高了数据读写速度,提高医生工作效率。

(3) 数据中心与现有信息系统集成方便。由于 Hadoop 架构的各个组件提供了多样化的编程开发接口,与现有临床信息系统的集成工作易于实现,可以实现无缝集成。

(4) 数据中心提供机器学习平台。基于 Hadoop 框架构建的数据中心具有云计算的能力,这使得对大数据的挖掘更加高效,同时,分布式文件系统提供的快速文件读写特性,也提高了数据挖掘的效率。

(5) 使数据中心易扩展。这主要分为两方面:当数据中心存储容量不足时,可通过为每一个节点的计算机添加存储容量即可;当数据中心计算能力不足时,为数据中心直接添加普通计算机即可,新添加计算机只需要简单地配置工作。

(6) 与现有数据中心相比,成本低。成本低主要体现在硬件和软件两方面:由于 Hadoop 整个框架是开源开发,其框架下的所有软件均免费,节省了构建数据中心的软件成本;另一方面,使用普通廉价 PC 构建数据中心,避免了在大型服务器上的高额投入。

8.3.3 医疗健康大数据业务架构

医疗健康大数据业务架构如图 8-2 所示。系统以居民个人健康卡为核心,搭建国家级人口健康信息平台、省级人口健康信息平台、地市级及县级人口健康信息平台,可以资源共享到社区医院。建立以电子病历数据、电子健康档案数据库、人口数据库等为基础数据库的数据中心,为医疗健康大数据分析作为决策依据。

通过建立、健全医院信息系统,加强以电子病历为核心的医院信息平台的建设,可以实现以下业务。

(1) 规范医院的财务管理,加强财务核算,改善医院收支状况。

(2) 加强医疗质量过程管理,减少医疗差错,提高医疗质量,保障医疗安全,增加病人的满意度。

(3) 优化整合医院的业务流程,提高工作效率;标准化医院的业务流程,提高工作质量。

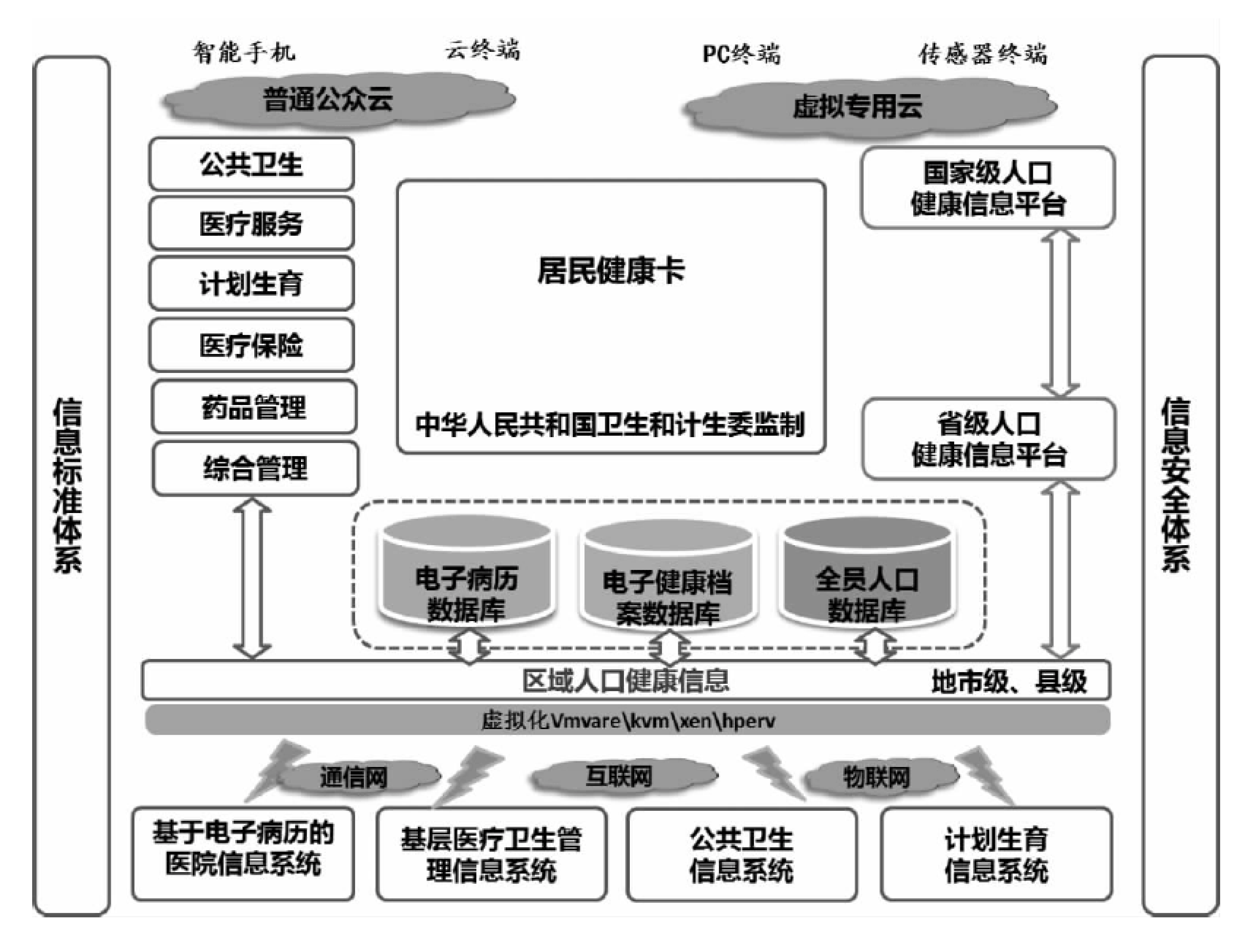


图 8-2　医疗健康业务架构图

(4) 加强各类业务数据的采集、传递、存储和使用管理,促进医院内信息共享,并为实现未来的医疗信息的区域共享打下基础。

8.3.4　医疗健康大数据技术架构

医疗健康大数据采用分布式数据库 HBase 和分布式文件系统 HDFS,如图 8-3 所示。省级数据中心和地区级与社区卫生信息进行数据交换。

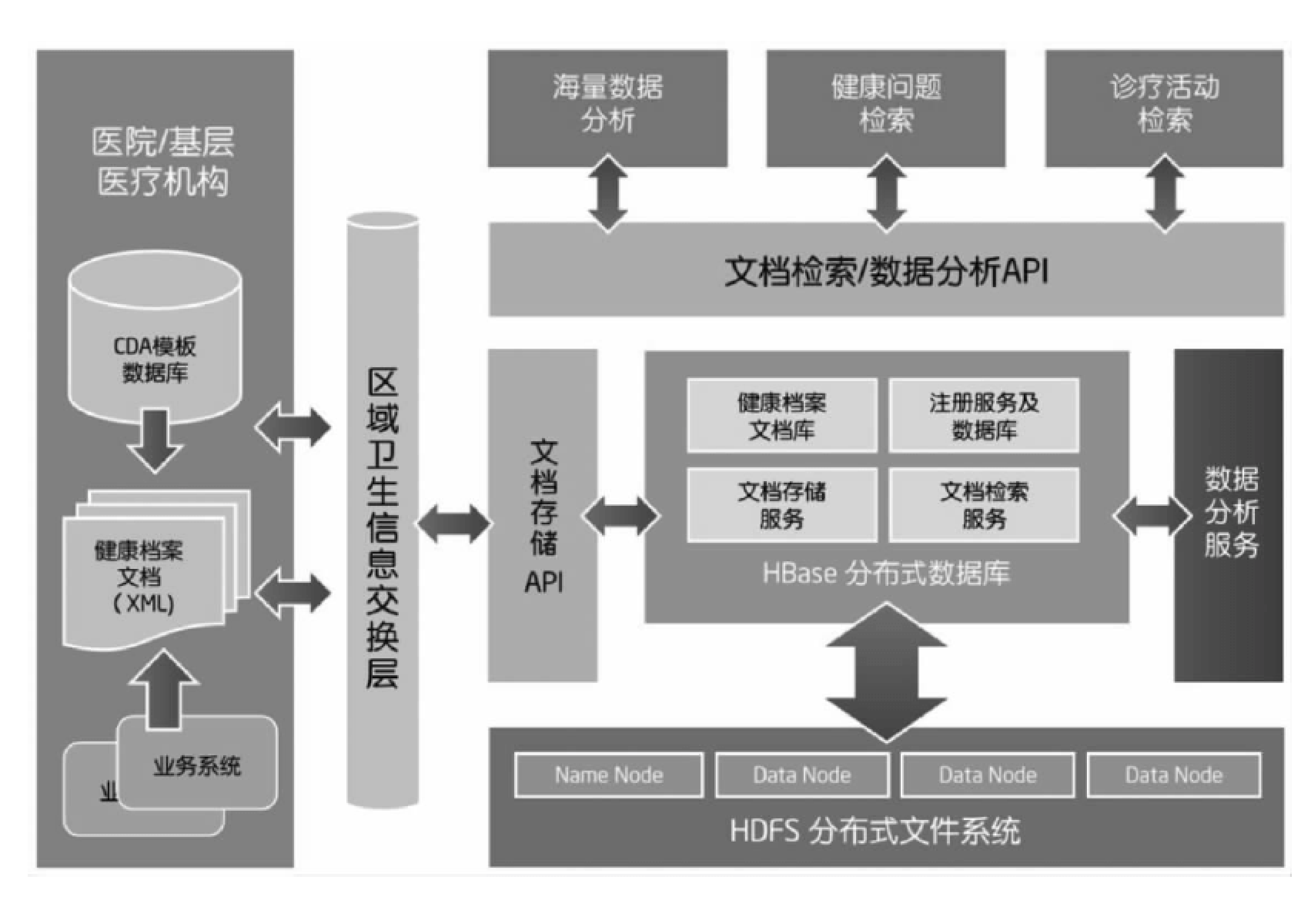


图 8-3　医疗健康技术架构图

8.3.5 医疗健康大数据网络架构

医疗健康网络架构如图 8-4 所示，以省级医疗健康数据中心为核心，进行资源集中，基于数据中心，构建 3 种基础设施能力，包括计算、存储以及网络的能力，面向三级医疗机构提供多样化的服务。

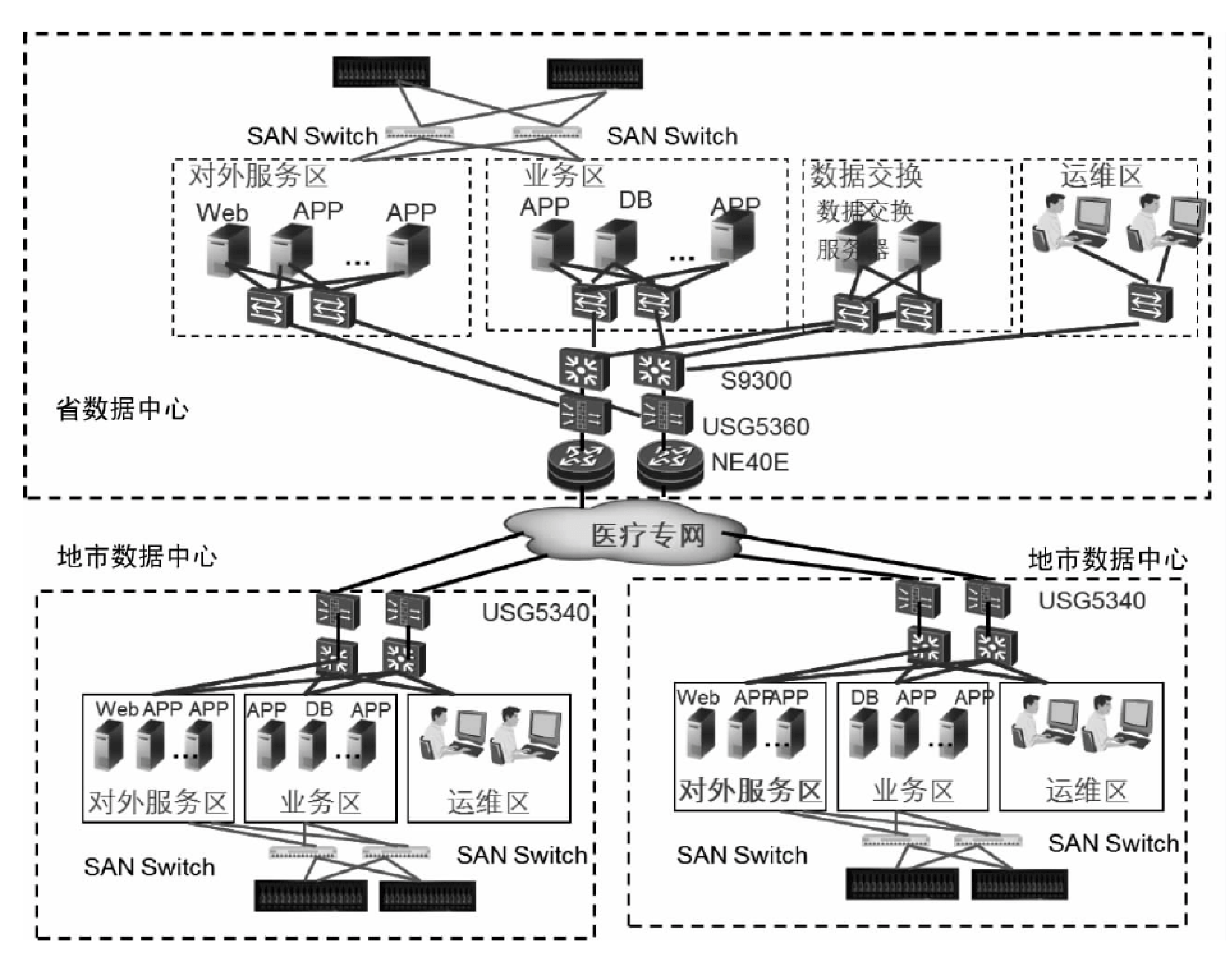


图 8-4　医疗健康网络架构图

医疗健康数据中心网络架构，构建云存储、云计算和云网络设备组成的云资源系统为医疗健康服务系统提供云计算的能力，使其具有强大的计算能力与共享服务能力。其中，云交换机和云路由器是一种新型可编程、可虚拟化、自适应和智能调度与资源共享构成的精简架构，即智能交换机和智能路由器，可以支持各种协议，实现接入普适化和控制智能化。

网络架构主要特点如下。

(1) 多层级管理，多集群设置：管理节点与业务节点分离，支持异构集群并存，管理节点本身支持分布式，使得管理节点性能超越单服务器性能局限。

(2) 超大容量集群规模：1024 服务器/每集群，同一集群内所有 VM 弹性分配存储空间，使得物理服务器的 VM 迁移及 HA 调度可在更大尺度内进行。

(3) 百 TB 容量无阻塞交换网：非组合的基于 CLOS 模型的多层级联式交换架构，支持 VM 间，以及 VM 与存储间的无组合、无丢包、低时延的性能。

(4) 基于目录服务的大层双交换网：引入网络目录服务，实现物理与逻辑 L2 地址的解耦，突破普通 L2 网络节点规模不超过 4096 节点的限制，支持大层双网络内所有 VM 的移动性。

8.4 医疗健康数据中心解决方案

在医院信息系统建设过程中，HIS 和 PACS 的医院信息系统数字化是两大重点，下面就对这两方面的需求分别进行分析。

随着医院信息化的不断深入，病人对医院的要求也越来越高，如果医院不能满足病人的合理需求，在一定程度上会造成医患纠纷等。因此，对医院信息系统的建设也提出了更高的要求，通常在一个医院信息化建设过程中，数据需求量最大的是 PACS，对数据安全要求最高的是医院核心系统 HIS。目前，PACS 已经成为现代医学放射学实践的基本技术和基础设施中重要的一部分，在临床诊断、医院科研等方面正发挥着极其重要的作用。

当前的 PACS 产品支持医学图像的全数字获取、转换、解释、存储和查阅。PACS 的发展也呈现出一个很大的特点：医院影像设备的发展使放射科图像数据激增，图像的数据量为存储容量带来了很大的挑战，数据需要进行分级存储和归档，同时，数据需要备份容灾和异构存储环境的现状也愈加突出，因此 PACS 需要一种可靠、灵活的大容量存储系统来满足其应用和发展。

存储系统的稳定性直接导致了 HIS 的业务连续性，当存储系统发生意外宕机时，整个医院运行将面临瘫痪，建设统一、安全、高可靠、分层的存储系统对医院信息系统的建设是至关重要的。

医院的 HIS 要对门诊、收费、药房管理和 OA 等系统提供服务，对存储空间的需求并不是很大，但对存储系统的性能和稳定性有着较高要求。

PACS 对患者大量的医疗和影像数据进行采集、存储、传输和处理。一个中等规模的三甲医院年平均的存储数据量至少在 2TB 以上，其中，PACS 的影像数据占据了 95% 以上。这样大数据量的资料存储、传输和处理对医院的网络平台、存储系统都提出了很高要求。

医疗行业对影像的要求非常苛刻，HIS/PACS 对存储系统自身的特点和要求，主要有以下几方面的特点。

(1) PACS 的影像图像主要是多媒体文档，并发访问量小，根据不同影像科室的特点，有的文件比较大，例如核磁阵列，有的比较小，例如 CT 等。HIS 核心通常都是数据库，例如 Oracle、DB2、SQL Server 等。

(2) 医疗 PACS 中的数据保存量大，数据量增长速度快，由于病人自身的情况，通常在前几个月医院会频繁调阅病人的医疗影像，后期很少调阅，但又不能对这些影像进行删除，因此，部分数据将作为归档数据，需要安全地保存和随时方便地调用，需采用分级存储策略。

(3) 随着医院数据量的激增，分级存储设计逐渐发展为在线、近线、离线的三级存储架构。

(4) 数据量大，达到海量存储。为了提高医院对病人服务的满意度，长时间等待调阅图像是病人无法忍耐的，诊断工作站和浏览工作站对在线图像检索速度的要求越来越高，甚至达到秒级。

(5) 部分影像资料用于科研和教学，重要性高，需要可靠、有效的容灾数据保护方案。

(6) PACS 和 HIS 数据各有特点，特别在存储容量、访问响应速度、访问频率、存储可扩展性等方面存在差异，需要分别考虑，有条件地进行分类存储。

(7) 随着医疗行业竞争日趋激烈，PACS 的建设需要投资的总成本较高，应该降低总拥

有成本,提高投资回报率。

(8) PACS 的设计需要具备高扩展性和灵活性,需要支持容量增长的高度可扩展架构和对异构存储环境的支持,以实现将来无缝扩容,而且不增加因扩容带来的管理开销。

医疗行业有着最为复杂的应用系统,每类应用对存储系统的需求千差万别,构建的存储系统需要涵盖多种应用的具体需求,除了需要考虑针对结构化数据(例如 HIS 的数据库数据)进行有效存储及保护外,同时还需要大量非结构化数据(例如 PACS 应用的图像、影像等数据)采用对象存储方式存放,并且需要进一步保证关键数据的备份和容灾。

8.4.1 医疗数据中心架构设计方案

随着信息化建设的进一步加强和深入,医疗卫生行业产生的数据量会越来越大,PB 级数据存储的时代会马上到来。这么大的海量数据如何管理和存储,如何能够最快地查询到需要的数据,如何进行关键数据的保护,如何进行存储优化,这些都是医疗卫生行业当前所面临的难题。存储可以帮助医疗卫生行业的客户有效解决海量数据环境下面临的各项挑战。

我国从事存储业务的厂商较多,存储产品种类比较齐全,覆盖面广,在海量数据处理方面具有丰富的产品,基于这些产品,针对对医疗行业存储需求的了解,本章介绍了医疗行业存储系统解决方案,用于满足医疗行业 HIS、PACS、OA 系统、ERM 系统的存储需求。

存储解决方案集先进的存储虚拟化技术、通用的硬件平台、优异的分布式文件系统和一体化的备份系统于一体,既能为 HIS、OA 等应用系统的数据库数据提供高可靠的结构化数据存储资源池,又能为 PACS、ERM 等应用系统的图片、病理文档的存储提供非结构化数据存储资源池,同时兼顾重要数据的备份和容灾。

医疗行业存储系统逻辑结构示意图如图 8-5 所示。

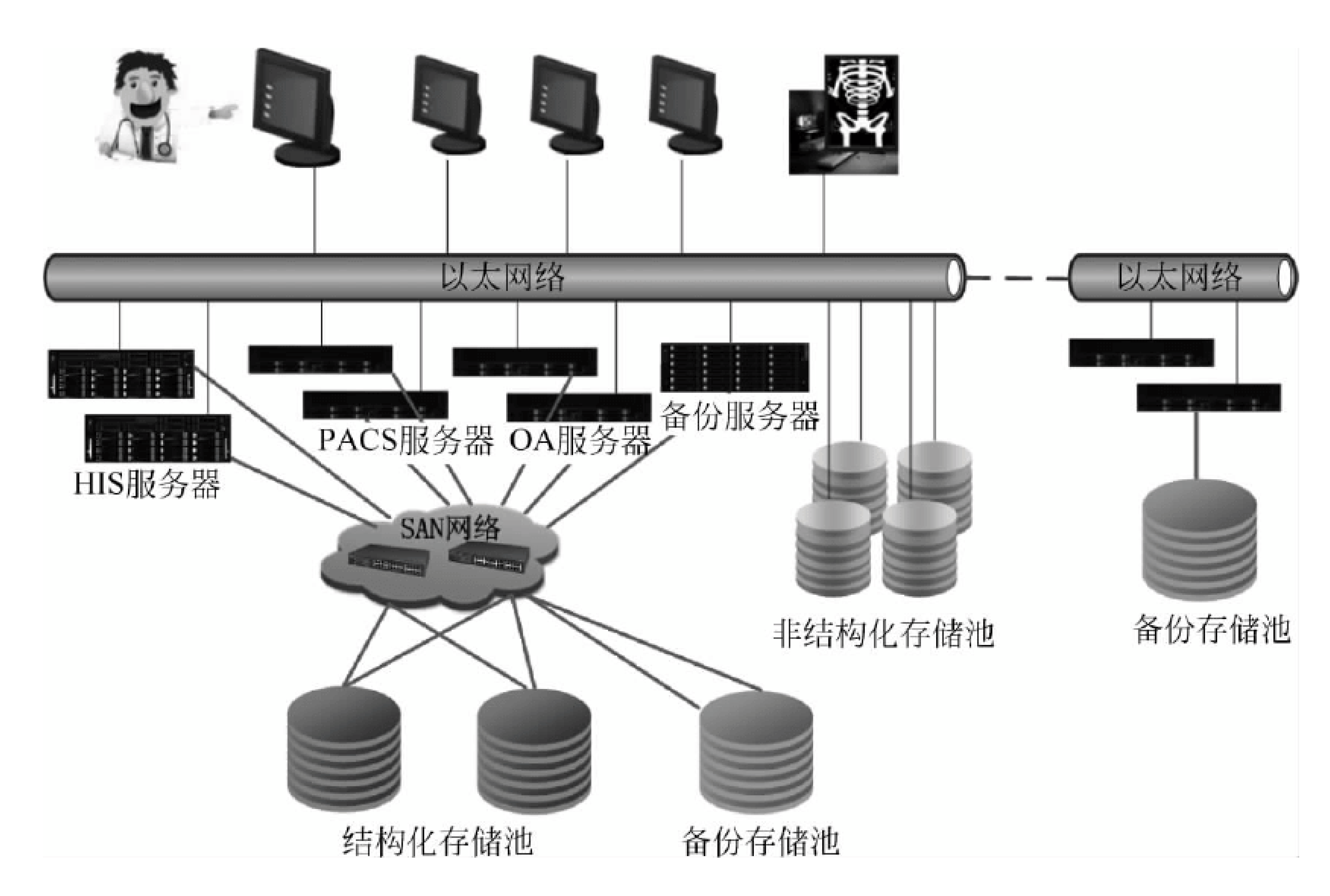


图 8-5 医疗健康存储架构图

为了有效解决医疗行业存储方面的需求，从 3 个方面设计存储方案如下。

(1) 结构化数据存储平台：块存储资源池面向用户的 HIS、OA 系统的数据库应用提供块设备存储空间，对存储系统的可靠性和性能要求比较高。结构化数据存储平台基于存储虚拟化技术构建，采用虚拟化控制器、FC 磁盘阵列和 FC 交换机组成一个全冗余架构的结构化数据存储平台，该平台可以实现关键数据的双写，保证业务平台不会因为任一存储部件或是存储单元出现故障而导致应用中断，具有较高的可靠性，同时存储设备之间的切换对上层应用透明，不需要人为干预。为了保证存储系统的高性能，底层设备采用最高端的 FC 磁盘阵列构建。

(2) 非结构化数据存储平台：非结构化数据存储平台为用户的 PACS、ERM 等系统的视频、图片、病历共享存储应用提供存储空间。该平台基于集群存储系统分布式存储系统构建。集群存储系统基于业界先进的集群技术、多副本技术、并行读写技术和 Scale-out 扩展技术构建，底层硬件全部采用商业标准单元，为用户提供全局单一命名空间。具有容量大、性能高、扩展方便、高可靠、易管理以及构建成本低等特点，解决了医疗行业的海量非结构化数据的存储问题。

(3) 一体化的备份容灾平台：为了降低和减少人为误操作以及自然灾害对关键数据造成的影响，公司为医疗用户提供了一体化的备份容灾平台。该平台基于 DBstor 集中备份系统构建，该系统集备份服务器、备份软件、备份存储空间于一体，使用方便，管理简单。支持异构平台的多种数据库系统以及文件的备份，支持本地数据备份和远程数据容灾，为医疗行业的各类数据库、重要文件提供高性能的保护。

8.4.2　集中存储解决方案

HIS 是医疗行业最为关键的生产系统，其数据类型主要为数据库数据。该系统对数据的可靠性要求很高，需要存储系统满足 7×24 小时高可靠运行的业务连续性要求，并且，随着就诊人数的增长，需要保证存储系统的性能和容量可以满足业务发展的需要，同时，为了避免因数据丢失引起的医疗纠纷，需要保证数据的安全性和可恢复性。

针对 HIS 的应用特点，我们认为 HIS 存储体系架构应具备以下特点。

(1) 采用高可靠的存储高可用体系架构，存储网络、存储设备均无单一故障点；

(2) 采用高 CPU 处理能力、高缓存性能、高可靠性、高稳定性的政府单位级存储系统；

(3) 采用数据备份技术进行数据保护。

针对医院 HIS 的应用特点和数据结构，基于存储虚拟化技术的存储高可用存储架构能够为其提供高可靠的存储服务。存储高可用系统采用数据双写技术，确保同一份数据在两套存储设备上各存一份，解决了长期困扰用户的存储设备单一故障点问题。

底层存储设备选用高性能、高可靠企业级 DS800-G20 FC 盘阵，该系统采用新一代高性能 Xeon 处理器与最新的 8Gb FC、6Gb SAS 接口技术，满足用户业务系统的性能需要，同时采用创新的 ACP(Automatic Cache Speed)技术，通过智能分析算法，可透明移动热点数据至高速存储空间(SSD)，可以显著提升 HIS 应用系统数据库的性能。

同时，为了进一步保证数据的安全性，本方案采用数据备份技术对关键数据进行备份和容灾保护(详见容灾备份系统内容)。

1. 高可用存储方案介绍

随着服务器高可用、网络高可用技术的发展越来越成熟,存储系统成为应用系统的单一故障点,虽然可以采用数据备份、磁盘阵列卷拷贝技术增加系统的可靠性,但是数据备份只能解决数据的逻辑错误,解决不了磁盘阵列的硬件故障,卷拷贝技术虽然可以解决硬件故障,但是该方案要求两套存储系统必须是同一厂商同一系列的具有卷拷贝功能的高端磁盘阵列,成本高,并且不能实现自动切换。

采用存储虚拟化控制器,上述问题得到更优化的解决。

2. 方案拓扑结构

如图 8-6 所示,结构化数据存储平台由两台存储虚拟化控制器、主磁盘阵列、备份磁盘阵列,以及冗余 FC SAN 网络所组成。两台存储虚拟化控制器之间通过光纤(FC)作为数据同步的心跳线。两台存储虚拟化控制器分别通过光纤接入到两套冗余的 FC SAN 网络中,从而实现对主、备磁盘阵列物理存储空间的接管,并为 HIS 高可用应用服务器集群提供虚拟的 VDisk 存储空间。

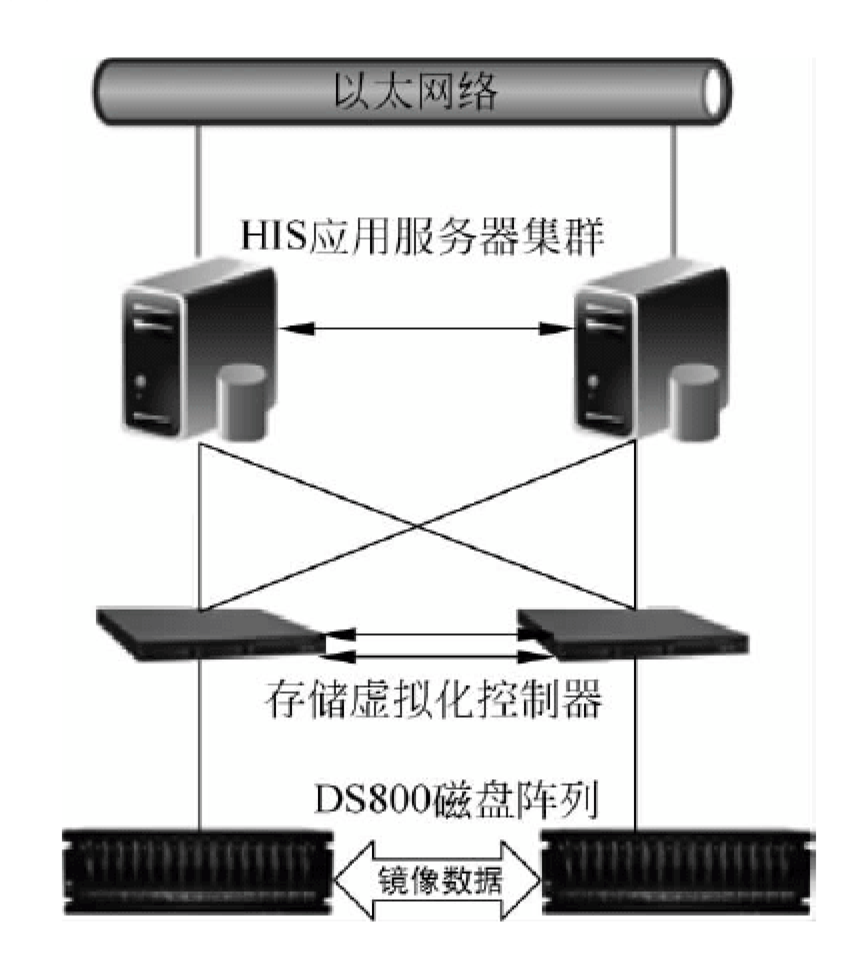


图 8-6 结构化数据存储平台拓扑架构

如图 8-7 所示,每当应用主机向 VDisk 中写入数据时,两台存储虚拟化控制器之间通过光纤,在两台设备之间进行同步镜像抄写。只有当数据成功被写入两台存储虚拟化控制器之后,才会返回 SCSI ACK 信号,通知主机操作成功。因此,所有写入的数据都实时地保存在了两个存储虚拟化控制器中,实现了数据在线热备保护。主、备存储虚拟化控制器选择在一个比较适合的时间,把保存在控制器中的数据写入所对应的主、备磁盘阵列中。

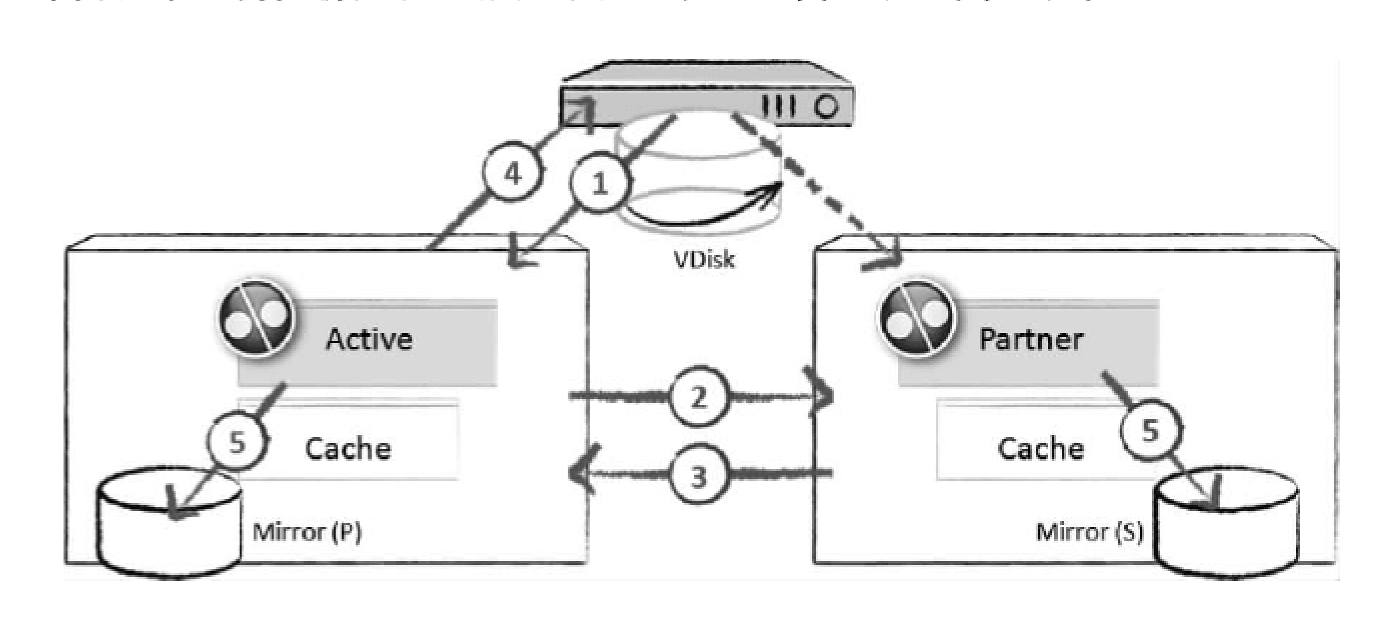


图 8-7 数据同步镜像

如图 8-8 所示,当主存储设备发生故障时(断电、端口故障、链路中断等),安装前端应用服务器中的多路径(MPIO)软件将自动进行存储路径切换(Auto Fail-Over),实时地把存储路径指向备份存储设备上。在此期间,应用服务器上的业务完全不会受到中断,保证了出现存储硬件故障情况下的应用业务连续性。

当主存储设备故障修复之后,MPIO 将自动把存储路径回切(Auto Fail-Back)到主存储设备上,同时,主机所在备份存储系统所做的数据变更,也会根据 I/O Update log,自动同步到主存储设备中,此过程无须人工干预,并且对应用主机的业务而言也是透明的。

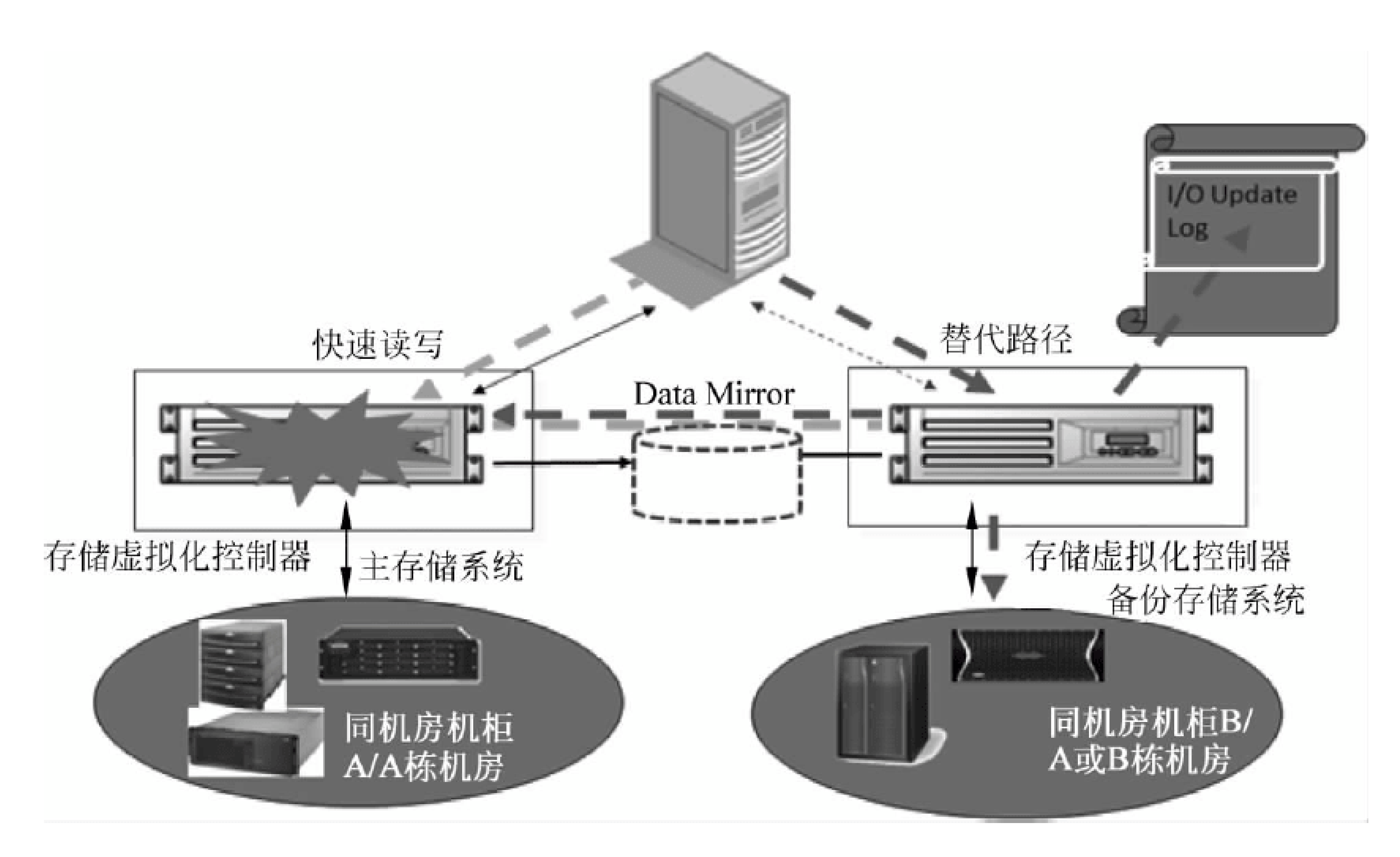


图 8-8　故障切换工作流程

存储虚拟化控制器支持各种主流应用主机集群软件，包括 Windows MSCS，RedHat Cluster，Rose HA 等，存储虚拟化控制器同时支持众多厂商的磁盘阵列产品。通过应用主机高可用与存储设备高可用的结合，可以形成完整的业务连续性保护方案。

为了尽可能保证应用系统的可靠性，建议用户为该 HIS 高可用系统配置两个机柜和两个 UPS。一个机柜内放置主存储虚拟化控制器、主磁盘阵列、一台 FC 交换机，并放置一台 UPS 为这些设备提供电源保护；另外一个机柜放置备份虚拟化控制器、备份磁盘阵列、一台 FC 交换机，也放置一台 UPS 为这些设备提供电源保护。这两个机柜可以放置在同一机房中，也可以放置在同一建筑的不同机房中。

3. 方案优势

HIS 存储高可用解决方案基于存储虚拟化技术，具有如下特色和优势。

(1) 该方案基于存储虚拟化技术实现，主要用于解决磁盘阵列单一系统故障问题。

(2) 该方案基于存储虚拟化控制器的卷镜像功能实现，具有硬件无关性，因此源端和容灾端磁盘阵列可以采用不同型号的磁盘阵列，并且磁盘阵列不要求具有快照和远程卷镜像高级功能，可以利用用户原有存储设备，最大程度地节省用户的投资成本。

(3) 与应用无关，支持所有类型的数据同步，包括文件数据、数据库数据、裸设备、应用配置文件、应用程序、库函数等。

(4) 支持同步和异步两种数据复制方式。同步方式用于本地或同城数据容灾方案，可以很好地保证数据的一致性；异步方式用于远程数据容灾方案，受数据复制线路影响，数据延迟大，不能保证数据的完全一致性。

(5) 为确保不存在任何单点故障，该方式会将两份镜像数据保存于不同的物理存储中，当其中源端存储设备发生故障时，应用服务器通过多路径存储技术，实时地将数据存储路径无缝地切换到备份端存储设备上。切换过程无须人工干预，并且应用不会中断，而未来在故障修复后，存储虚拟化控制器会自动地将两份数据进行同步。

(6) 故障切换时间极短，单位为秒级。

8.4.3 PACS 数据存储方案

PACS 是医院信息系统重要的组成部分，伴随着医院规模的不断扩大，PACS 的影像文件数量增长迅速，容量越来越大，且 PACS 的数据通常会保存长达 15 年甚至更长时间，其中有少部分的数据需要经常使用，绝大部分的数据属于历史数据，一般情况下这些历史影响数据很少被调用甚至在病患痊愈后再也不会被调用。这就造成在 PACS 中存在海量的历史静态数据。面对高达几百 TB 的医学影像资料，常规的数据存储已经不能满足 PACS 对数据的管理要求，这就对 PACS 的影像存储和管理提出了新的要求。

针对 PACS 的应用特点，我们认为 PACS 存储体系架构应具备以下特点。

(1) 采用专用的并行文件存储系统，管理日益庞大的海量影像文件；

(2) 采用在线-近线的分级存储体系架构；

(3) 采用集中式管理系统；

(4) 采用数据备份技术，对在线影像数据进行保护。

针对医院 PACS 的应用特点和数据结构，并行存储系统专有的文件系统能够更有效地管理 PACS 千万级的影像文件，并提供高速的数据访问能力，提供有效的数据共享能力。

PACS 应用服务器通过 IP 网络，利用文件协议连接到并行存储系统，实现 PACS 影像的在线存储。

并行存储系统内部将最新的影像数据保存在高速 SAS 磁盘上，将近期不频繁使用的历史影像保存在存储系统的 SATA 磁盘上，并确保历史数据能够实时地被业务系统访问。所有的数据由并行存储系统进行集中存储和管理。

同时对关键的在线影像数据利用数据备份技术进行数据保护。

1. 并行存储方案介绍

由于医院 PACS 中大量医学影像文件具有容量要求，并有数据保护要求高、连续性要求高和需要分级存储的特点，通常采用集群并行存储系统实现海量医学影像的集中存储和快速文件读取，并同时利用 DBstor 实现关键数据的备份。

集群并行存储系统基于开放式的存储架构，基层采用集群并行文件系统，将多台物理存储设备(这些物理设备可以是通用的存储服务器，也可以是磁盘阵列)的存储空间虚拟成一个具有统一访问接口和管理界面的存储池(也叫统一命名空间)。用户的数据按照一定的负载均衡策略，条带化地分布到后台的多套存储设备上，从而能够实现数据的并行读写以获得更高的并发访问性能，充分利用多台存储设备的性能和更大的存储容量，并有效地提高存储空间利用率，同时基于集群并行文件系统的数据迁移功能，可以实现实时和历史影响数据的分层存储，并且所有的存储设备可以实现统一的管理和监控，大大减轻了管理工作负担。

集群并行存储系统汇集了海量数据处理方面的核心技术，从架构上彻底消除了传统存储系统的瓶颈，能够满足高带宽和高并发的海量文件存取的需求，为用户带来前所未有的存储性能体验。

2. 集群存储拓扑结构

PACS 集群存储拓扑结构如图 8-9 所示。

存储系统包括管理控制器 MGR、索引控制器 oPara、数据控制器 oStor。其中，管理控

制器通过管理网络监控系统的各个模块的状态，提供统一的控制管理界面，实现存储系统的集中部署和监控，一套只需要配置一台管理控制器即可；索引控制器用于管理存储系统的所有索引数据和命名空间，对外提供单一的全局映像，一套集群存储系统一般至少需要两个索引控制器，两个索引控制器以 Active-Active 高可靠模式运行，一个控制器出现问题，不会影响存储系统的正常运行，索引控制器可以按需以成对的方式进行扩展；数据控制器用于提供文件数据 I/O 通道和实际的数据存储空间，并实现存取的动作，数据控制器根据用户实际的带宽以及容量需要进行配置，并可以按需进行动态添加，I/O 通道具有千兆、万兆和 IB 多种选择，并且集群存储系统支持多个 I/O 通道的冗余和负载均衡，为了保证数据的高可靠性，数据存储采用了多副本的数据保护技术。

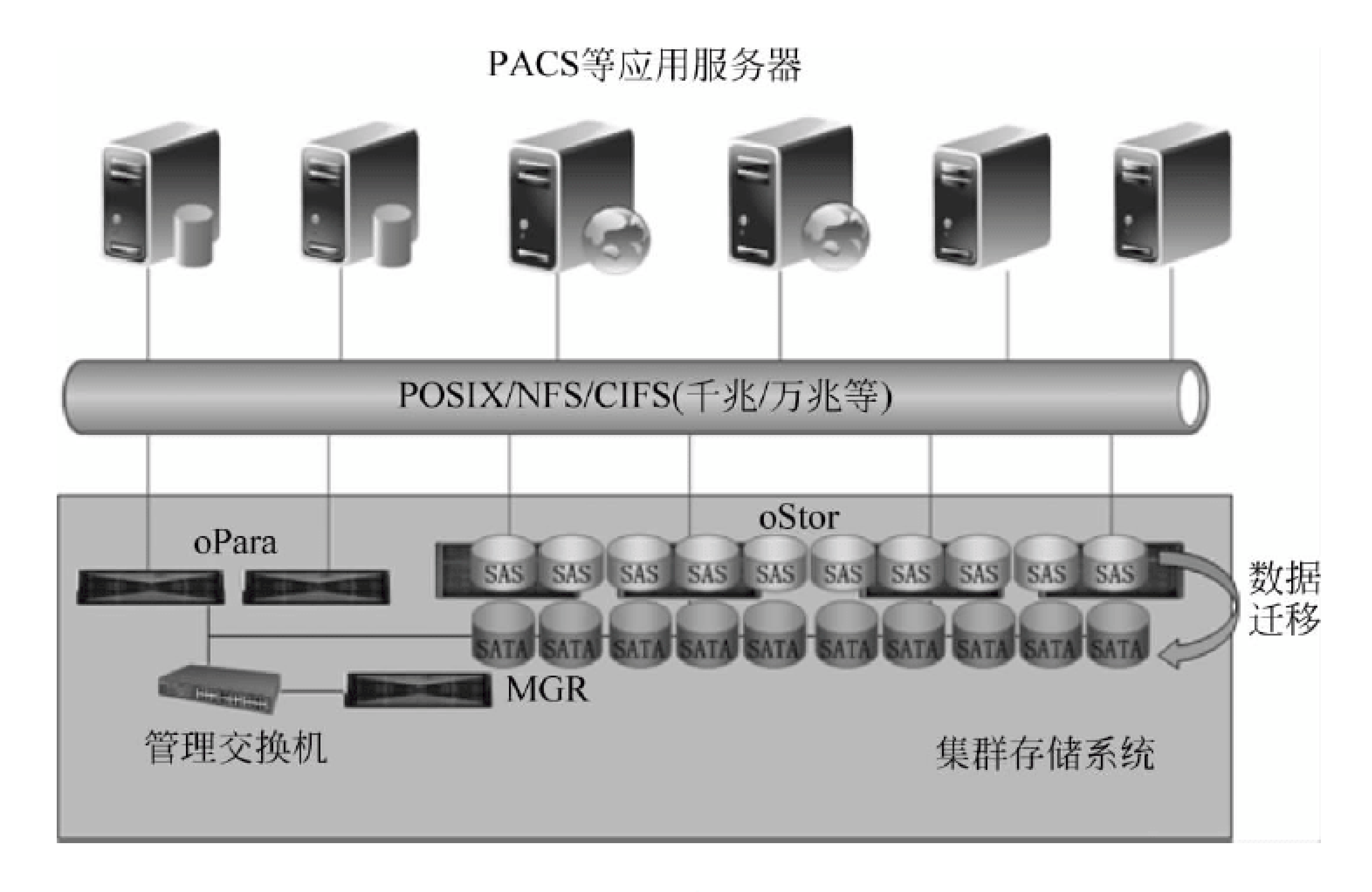


图 8-9　集群存储拓扑结构

oStor 可同时为实时影像数据提供高性能在线 SAS 存储空间以及为历史影响数据提供大容量近线 SATA 存储空间。集群存储系统支持基于策略的数据迁移功能，在线数据会在规定的时间内迁移到近线存储空间里去。由于每台 oStor 支持灵活混合配置 SAS 在线存储空间和 SATA 近线存储空间，数据的迁移在 oStor 内部即可完成，大大减轻了数据网络的负载，不影响前端应用存储访问的性能。

PACS、ERM 等前端的应用服务器可以通过两种方式访问集群存储系统：一种方式是通过应用系统提供的专有 Linux、Windows 客户端软件，这种方式没有额外的开销，性能较好；另一种方式是通过集群存储系统 NAS 模块提供的通用 NFS、CIFS 协议，这种方式支持的操作类型更为丰富，使用也更为简单，但是性能与第一种方式相比有所降低。

3. 方案优势

集群存储系统采用了代表存储技术、网络通信技术以及数据管理技术发展方向的并行体系架构，是一款面向海量非结构化数据处理、拥有自主知识产权的高端存储系统。

它具有如下特色和优势。

（1）单一命名存储空间，集中化共享虚拟存储池。

PACS 集群存储系统可以智能地将数据存放到存储系统的数据节点上，创建一个集中化的共享虚拟存储池，提供全局单一的命名空间。目前业界有很多存储系统也声明支持 PB

级的单一命名空间，但是底层无一不是通过将若干卷挂载在同一个根目录下来形成的大容量统一命名空间，其效率和出现存储热点时的性能，将会大大低于将上PB级别的存储空间置于同一个文件系统下管理的统一命名空间。

可以带来如下好处：①提高存储空间的利用率，高达90%；②简化海量数据管理的复杂性，用户可以直接对虚拟资源池进行管理和控制，无须考虑存储设备的布局方式；③超越传统存储架构容量和性能的极限。

（2）高性能并行存储系统，支持并发I/O读写，提供高达数百GB/s的聚合带宽。

衡量一个存储架构的优劣，无外乎从读写两个方面来看，集群存储系统的聚合性能可随着数据控制器节点的增加而增加，根据实际测试结果，集群存储系统数据控制器节点每个插两块双口千兆以太网卡，提供4个数据传输通道，单节点可以提供高达150MB/s的写带宽和360MB/s的读带宽。集群存储系统的聚合带宽，可以用每节点带宽乘以节点数来计算，系统性能可实现线性增长。部署于深圳云计算中心的、系统总容量16PB的集群存储系统可提供高达一百多GB/s的聚合带宽。

集群并行存储系统是如何达到超高性能的呢？通过如图8-10所示的存储系统的读写机制可以来分析说明。

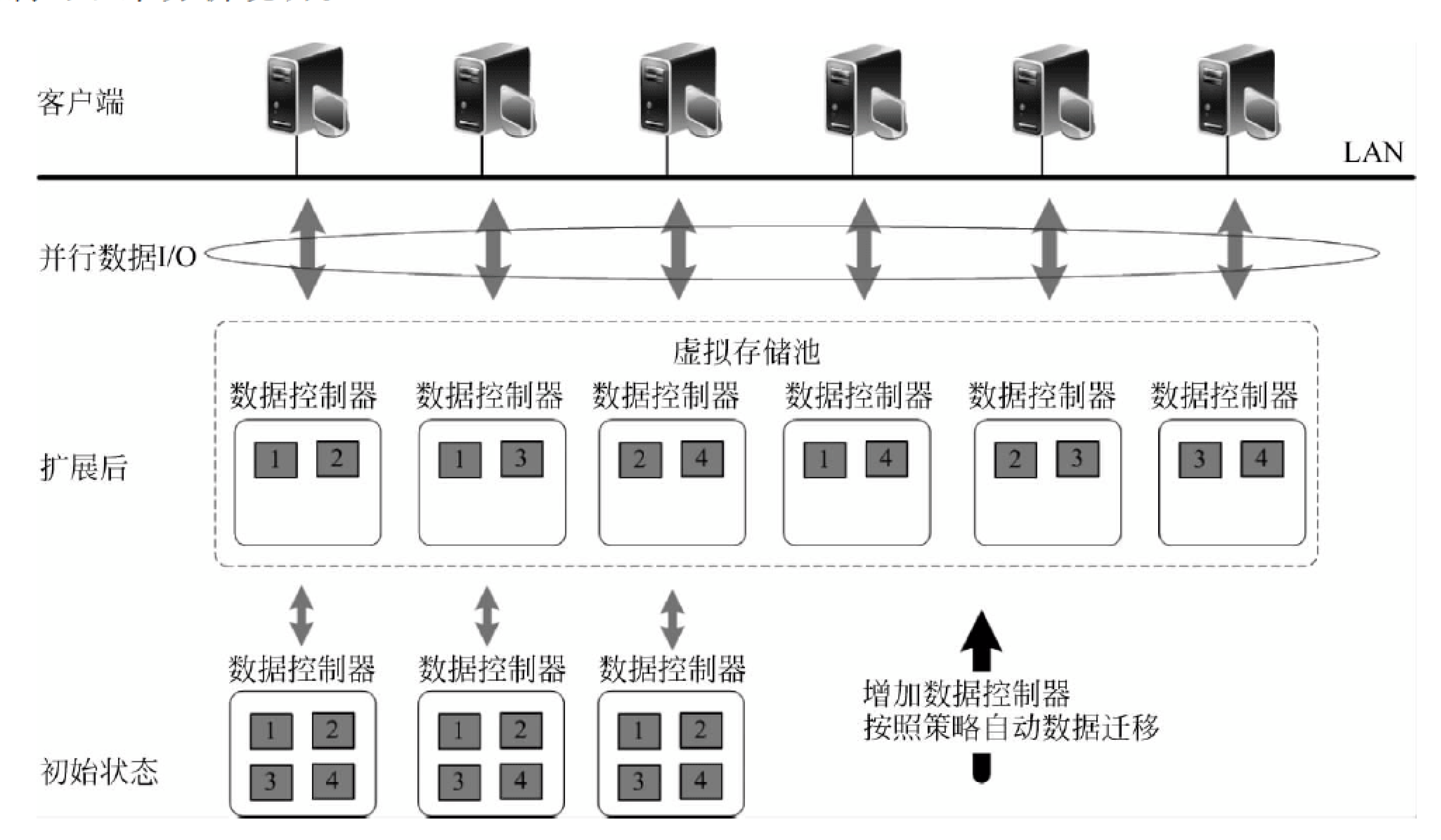


图8-10　集群存储读写机制

从客户端发起读写请求，通过核心交换机，向集群存储系统发起读写请求。以两个索引控制器和8个数据控制器为例，索引控制器首先会接收此次读写请求，并通过分析数据控制器的状态来确定将文件如何分割以及写到哪些数据控制器上，然后将此信息反馈给客户端，客户端通过读取此信息，利用与数据控制器相连的数据通道并发地将文件块写入到对应的节点磁盘上。反过来看读一个文件，这个非常好理解，就是写的一个反过程。形象地打个比方，传统架构数据要写到磁盘上或从磁盘上读取数据，相当于一个人搬8个箱子，而并行存储系统是8个人搬8个箱子，效率和速度大大提高，这种并行架构决定了系统读写的性能比一般的存储性能高很多。

(3) 存储系统自动实现数据分层,有效地提升读写性能、降低构建成本。

集群存储系统的存储介质可以根据实际需要选择高性能 SSD、FC、SAS 硬盘或是大容量高性价比 SATA 硬盘,这种灵活的构建方式可以极为方便地为 PACS 应用构建高性能存储区和大容量低速存储区,高性能存储区可以用来存储用户最新的一些访问频繁的医学影像素材及病历信息,低速存储区用来存储用户海量的历史医学数据和病历信息,集群存储系统支持用户采用自动或是手动的方式实现数据在这两个区的迁移,如图 8-11 所示。

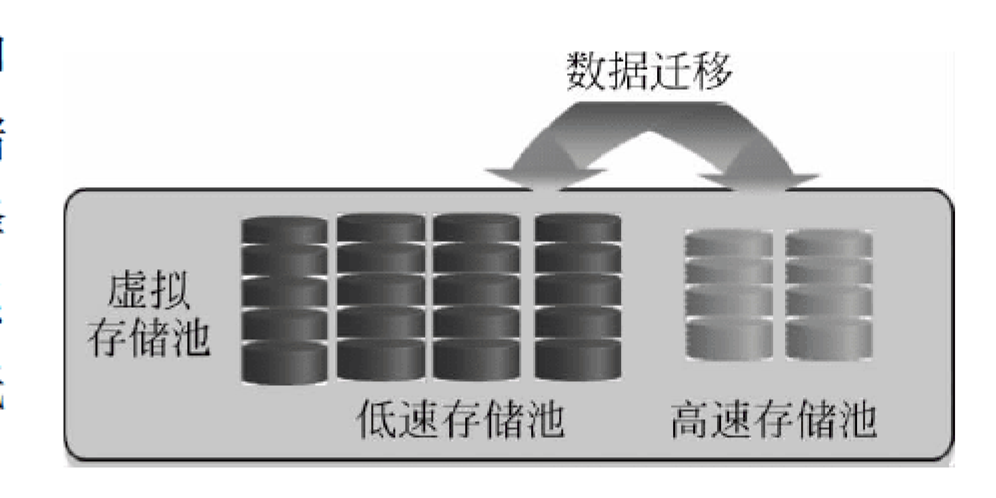


图 8-11　数据迁移示意图

集群存储系统这种灵活的分层存储架构和对数据迁移功能的支持,使得 PACS、ERM 存储系统不需要额外的数据迁移管理系统,有效地降低了存储系统的构建复杂度,同时提高了存储系统的可靠性以及查询的效率,在最大程度上降低了用户的投资成本以及使用和管理成本。

(4) Scale-out 扩展方式,性能和容量随数据控制器数量的增加而线性增长。

Scale-up 向更强大的 CPU、内存、通道及其他设备扩展,而 Scale-out 则是通过一定的技术将一个个独立的低成本存储节点聚合成一个大而强的系统。对于用户来讲,Scale-up 架构的存储系统其设备处理能力上限在购买时已经确定,如果需求超过上限,只能重新购买更高性能的设备;而 Scale-out 架构的系统可以采用搭积木的方式,通过把成百上千台个体聚合起来,来满足不断增长的存储容量和性能的需求。

集群存储系统就是一款基于 Scale-out 架构的高端存储系统。

由于采用了 Scale-out 横向动态扩展技术,集群存储系统可以为用户提供如下好处:①打破了传统架构的扩展限制,容量可扩展到 EB 级,同时可以保证存储系统容量和性能的均衡;②避免由于用户需求的不断增长而带来的数据迁移和应用中断的问题,可随业务需求变化而动态调整资源,带宽、处理能力和存储容量都可以实时调整和扩展;③存储容量和聚合带宽随着数据控制器的增加动态线性扩展;④所有扩容操作均可以在线进行,无须中断应用的正常执行。

(5) 集群存储系统采用多副本、全冗余架构等多种先进的数据保护机制确保整套存储系统安全可靠,无单点故障,提供始终如一的高可用性。

传统的 NAS 和 SAN 存储架构都存在固有的单点故障,容易形成"数据孤岛",一旦 NAS 头或者磁盘阵列机头出现问题,发生数据不可用的危险就会大增。此外,传统的 RAID 技术,包括 RAID4、RAID5,在过去很长一段时期中都能基本满足需求,提供单个磁盘驱动器发生故障时的数据保护。但是现在都采用大容量磁盘驱动器,发生第二块、第三块磁盘故障的概率大增。面密度以将近 100%的复合年增长率在不断加大,但是磁盘驱动器的可靠性和性能并没有同步提高,而且由于大容量磁盘需要更长的 RAID 重建时间,极大地增加了同时发生几个磁盘故障的可能性,数据丢失的风险不可小视。再者,传统存储系统在发现和处理故障硬件部件问题时,都是被动反应,而非预先应对。因为不具备有预知功能的智能软件,不能预判什么时候会发生故障,传统存储系统将用户的数据置于危险之中。而集群存储系统由于采用了全冗余架构、数据多副本技术以及高效数据容错重构技术和故障自动恢复机制,系统可靠性和数据安全性非常高。

集群存储系统可以提供针对系统级和文件数据级的两个级别容错。

(1) 系统级容错。

数据读取和传输过程中,当某一个模块(可以是索引控制器、数据控制器或者交换机)发生问题宕掉了,通过系统级容错,冗余模块可以接替问题模块继续工作,系统仍是可用的,数据仍然是安全的、完整的,用户端感觉不到任何变化。整套系统没有单点故障。

(2) 文件数据级容错。

集群存储系统通过条带化技术将文件分块存储在多个数据控制器上,其中每一个分块都会有两个以上的副本存放在不同的数据控制器上。当客户端读取的数据块所在的数据控制器无法访问时,依然可以通过访问该数据块副本所在的数据控制器来读取数据,同时系统将会自动地在另一台可用数据控制器上生成此数据块的新副本。这种容错机制可以保证只要系统中剩余空间的容量大于损失的硬件中所存储数据的容量,系统即可自动进行数据恢复。同时,由于每一个存储设备上的数据所对应的另一个副本是分布在其他所有的存储设备上的,存储系统数据的恢复重构过程是一个多到多的数据复制过程,其恢复速度大大高于传统的存储系统,保证了业务的连续性和数据的安全性。

集群存储系统恢复重构 1TB 的数据只需半个小时左右,而传统的基于 RAID 技术的存储系统,即使是高端磁盘阵列,重构 1TB 的数据都需要十几个小时。

(1) 使用方式丰富。

集群存储系统为用户提供了丰富的使用方式。集群存储系统为追求极致性能的客户提供私有 Linux、Windows 接口应用模式,该应用模式需要在客户端安装提供的客户端软件,客户端软件不用修改和编译操作系统内核,这种直接访问方式没有额外的开销,因此具有最好的性能。

还为用户提供间接应用模式,在该应用模式下,集群 NAS 模块为用户提供 NFS、CIFS 标准访问协议,用户的前端应用服务器通过这些标准访问协议访问后端存储系统,这种方式对客户端没有影响,使用最简单。

(2) 管理方便。

集群存储系统集成图形化的并行存储管理软件系统,实现存储系统的统一管理和监控,有效地减轻管理工作负担。

并行存储管理系统是专为系列产品开发的统一监控管理平台,提供系统配置、客户端管理、性能优化、监控告警等功能,直观易懂的中文图形化界面方便用户实时监控系统的软硬件状态和性能,简化安装和维护过程,提高管理效率。

集群并行存储管理系统主要提供以下功能。

(1) 管理维护:提供服务启动停止和节点上线、离线功能。

(2) 文件系统管理:提供文件系统查询、文件系统创建、文件系统删除功能。

(3) 客户端管理:提供客户端查询、增加删除客户端以及修改客户端功能。

(4) 安装配置:为管理员提供系统的安装卸载、节点扩容删除、系统升级、数据删除、数据磁盘增加删除、配置修改、配置备份恢复功能。

(5) 监控管理:提供索引控制器、数据控制器,以及并行文件系统和整个存储系统的监控。

8.4.4 容灾备份解决方案

1. 一体化备份容灾方案介绍

考虑到大范围内灾难或故障发生的可能性，为了保障数据安全，利用现有存储设备资源为大量的主流平台用户制定完备的备份和容灾方案，构建简单、经济、可靠的备份及容灾系统，增强系统的抗灾能力，最大限度地减少损失有着十分重要的意义。

备份容灾系统是应用系统的补充，起到将应用系统中的数据（比如文件系统当中的文件、数据库中表的数据）形成副本，最终存放到适当存储介质（比如磁盘阵列、虚拟带库、磁带库等）当中，在应用系统数据损坏或者应用系统本身出现问题需要进行重建时，数据的副本为重建提供完整的数据来源，从而为应用系统提供最后一道安全防线。

容灾备份存储系统，通过采用两级方式对医疗行业的关键数据进行保护。数据在本地进行备份，然后远程保留一份，实现远程容灾。容灾备份存储系统通过一个统一的管理界面，对所有关键数据统一管理，实现数据保护，保证用户的业务连续性。

2. 容灾备份拓扑结构

备份容灾系统拓扑图如图 8-12 所示。备份方案首先使用 DBstor 设置合适的备份策略在本地进行备份。备份的数据类型可以是 HIS、OA 等应用中的数据库，也可以是 PACS、ERM 等应用中的图片、病历档案等文件。备份网络可以通过以太网络，有条件的用户也可以选择具有更高带宽的 Lan-free 备份方式，通过 FC SAN 网络把数据库数据直接复制到 DBstor 的备份空间里。本地局域网的带宽较大，可适当加大备份的频率。

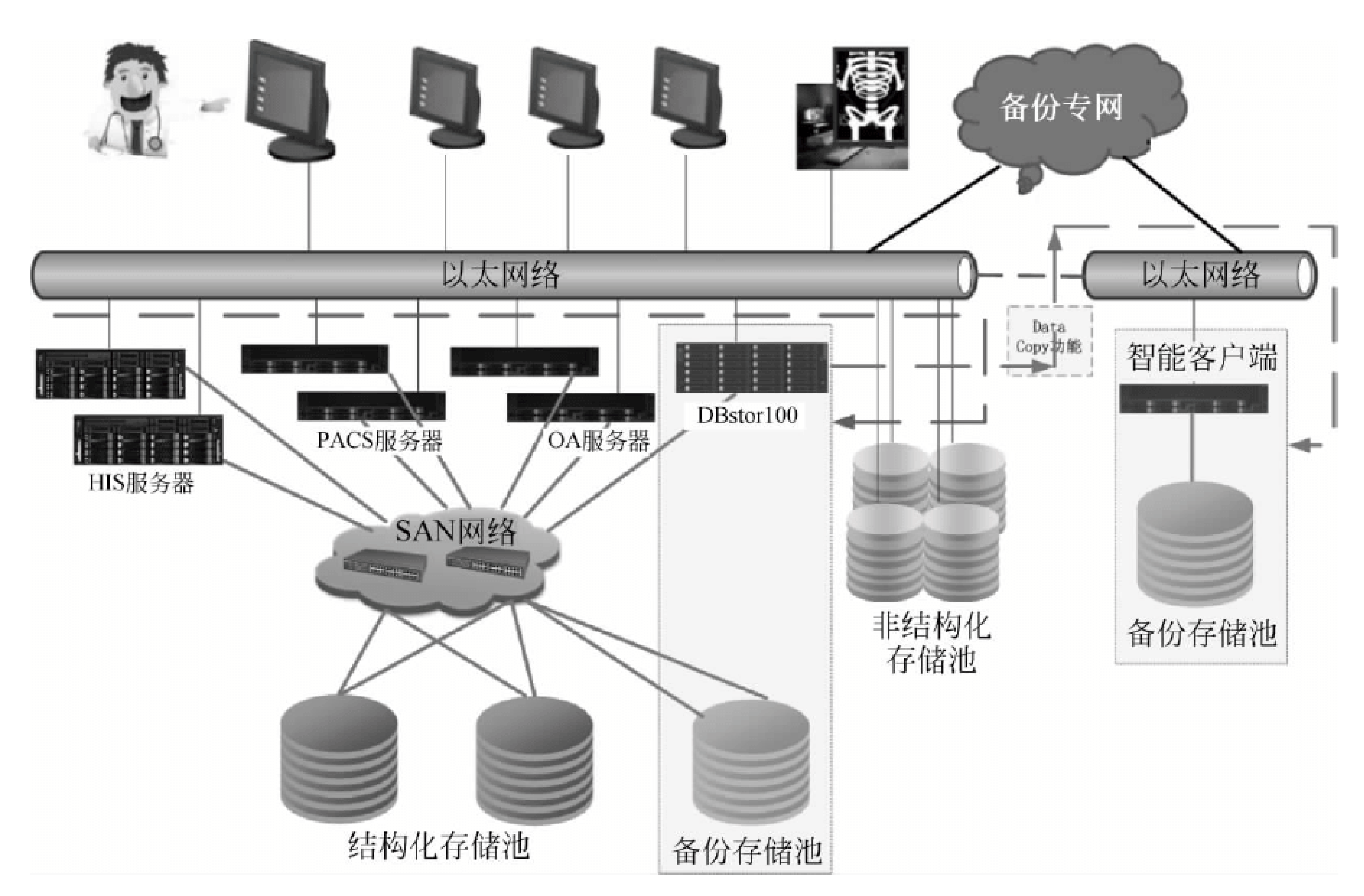


图 8-12 容灾备份系统拓扑结构

容灾方案需要在远程的容灾中心配置一台容灾服务器和一个磁盘阵列，容灾服务器上配置 DBstor 的一个智能客户端模块，利用 DBstor 的 Datacopy 功能将本地备份的数据复制到远程智能客户端所管理的磁盘阵列里。备份网络需要租用运营商带宽或建设专网，并要

根据网络状况和复制的数据量设置合适的时间点和策略。

DBstor 自身可以配置高达 48TB 的 VTL 或是带重复数据删除的 Smart Disk 备份存储空间，如果备份空间不够，还可以通过挂接磁盘阵列的方式进行备份空间的扩展。

如果应用系统出现人为误操作或是不可恢复的硬件故障所导致的数据错误，可以利用本地的备份数据进行恢复；如果本地的生产数据和备份数据因为自然灾害等原因全部出现故障，可以利用复制到容灾中心的数据恢复到用户的备用生产系统，接管客户应用。

3. 方案优势

方案基于数据备份技术实现，主要用于解决高端容灾(实时数据保护)不能解决的问题，如人为误操作、恶意性操作等。这类操作，计算机系统是不能区分的，一旦执行，将造成数据中心、灾备中心同时修改；对于数据库系统，在日志方式下，可以通过回滚方式修改，对于文件系统、操作系统等其他配置信息是不能回滚的，将造成毁灭性的结果。因此建设高端容灾系统的前提，是一定要做好本地系统的备份，这是容灾技术的基本要求。

软硬一体的备份容灾方案具有如下特色和优势。

软件、硬件一体化的备份容灾系统，并且备份系统同时又能升级为容灾系统，实施、使用、维护简单，大大减轻了用户的工作负担和人力资源。投入 DBstor 独有集成 VTL 模块，不需要单独 VTL 设备的支撑，减少了用户的投入成本，简化了管理工作。同时由于 DBstor 可以虚拟任意多的驱动器，可以实现多台数据库同时备份，拥有较高的备份频率，实现 RPO 很小；支持 LAN-Free 备份，可以得到很高的备份和恢复速度。DBstor 还可以提供性价比最好的 SmartDisk 备份存储介质，具备重复数据删除技术，凭借其强大的基于软件的字节级可变数据块去重技术，可以减少存储成本。DBstor 具有良好的兼容性，支持 Windows、Linux、AIX、HP-Unix、VMware 等各种异构客户端，同时支持 Oracle、RAC(Linux 版本)、SQL Server、Sybase、MySQL 等数据库的备份简单、自动化，无需脚本。将文件备份、数据库备份、操作系统备份集中在一个统一的管理界面下，对各种介质的管理、各种备份设备的管理、策略的管理，集中在一个统一的软件中；支持数据库在线联机备份，定制策略和恢复过程纯图形界面，不需要编辑脚本。

部分点采用光纤带库，LAN-Free 备份，可以得到很高的备份和恢复速度。基于图形界面的集中化数据备份方式，中文操作界面，便于用户使用、维护。尤其是 ReportManager，可以通过基于颜色的图形界面，发现备份的问题，便于多点集中监控。自动通知功能非常方便，可以通过邮件发送报告。对于将来增加的数据库服务器、应用等服务器，只需在新增相应主机上安装相关的客户端软件、SmartClient 软件(如果接入到 SAN 中)、相应数据库接口软件包(如果运行数据库)即可，便于备份的扩充。

8.5 医疗健康大数据分析

伴随着中国医疗卫生服务的信息化进程推进，将产生大量的数据。这些数据主要来源于医疗业务活动、健康体检、公共卫生等 9 项医疗卫生服务。数据内容包括来自医院的大量电子病历、区域卫生信息平台采集的居民健康档案等。其中大量充斥着非结构化/半结构化的数据，包括图像、Office 文档，以及 XML 结构文档等。医疗大数据的应用，关键是整合所有可能得到的这些数据，为机构和政策制定者提供找到如何刺激经济并降低共享数据技术

门槛的途径。

8.5.1 医疗实体对象建模分析

我国医疗卫生行业涉及的数据实体对象种类非常多，包括医疗机构-科室-医生(门诊、住院)、大众群体-患者、医疗管理部门-卫生局-疾控中心-医保中心-发展改革委员会-中医药管理局、医药管理部门-药品监督局、医药研发-医药生产-医药经营-药品(处方药、ODC 药)、医疗器械研发-医疗器械生产-医疗器械经营-医疗器械、商业医疗保险公司、体检中心-体检医生、APP 服务等。

如图 8-13 所示是数据实体对象建模示意图。

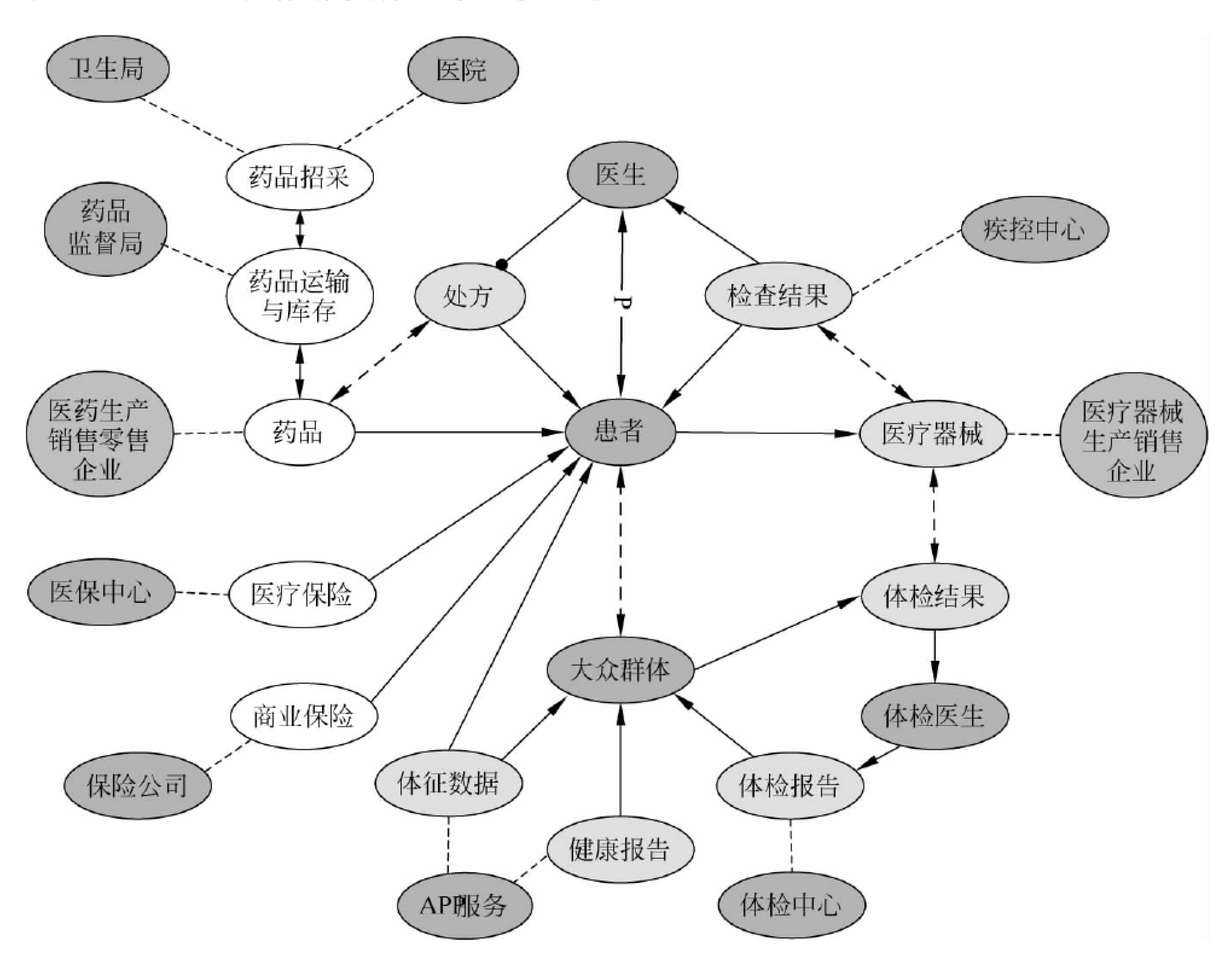


图 8-13 数据实体对象建模图

8.5.2 医疗个人健康档案建模分析

2010 年年底，原卫生部完成了“十二五”卫生信息化建设工程规划编制工作，初步确定了我国卫生信息化建设路线图，简称“3521-2 工程”，即建设国家级、省级和地市级 3 级卫生信息平台，加强公共卫生、医疗服务、新农合、基本药物制度、综合管理等 5 项业务应用，建设健康档案和电子病历两个基础数据库和一个专用网络建设，进行医疗卫生信息标准化体系和安全保障体系两个体系建设。

2013 年 11 月，卫生部和计划生育委员会合并后，信息化建设工程规划的顶层设计规划又调整为“4631-2 工程”，其中，“4”代表 4 级卫生信息平台，分别是：国家级人口健康管理平

台，省级人口健康信息平台、地市级人口健康区域信息平台及区县级人口健康区域信息平台；"6"代表6项业务应用，分别是：公共卫生、医疗服务、医疗保障、药品管理、计划生育、综合管理；"3"代表3个基础数据库，分别是：电子健康档案数据库、电子病历数据库和全员人口个案数据库；"1"代表一个融合网络，即人口健康统一网络；最后一个"2"是人口健康信息标准体系和信息安全防护体系。依托中西医协同公共卫生信息系统、基层医疗卫生管理信息系统、医疗健康公共服务系统打造全方位、立体化的国家卫生计生资源体系。卫生和计划生育委员会规划的三大基础数据库相互关系和包括的主要数据如图8-14所示。

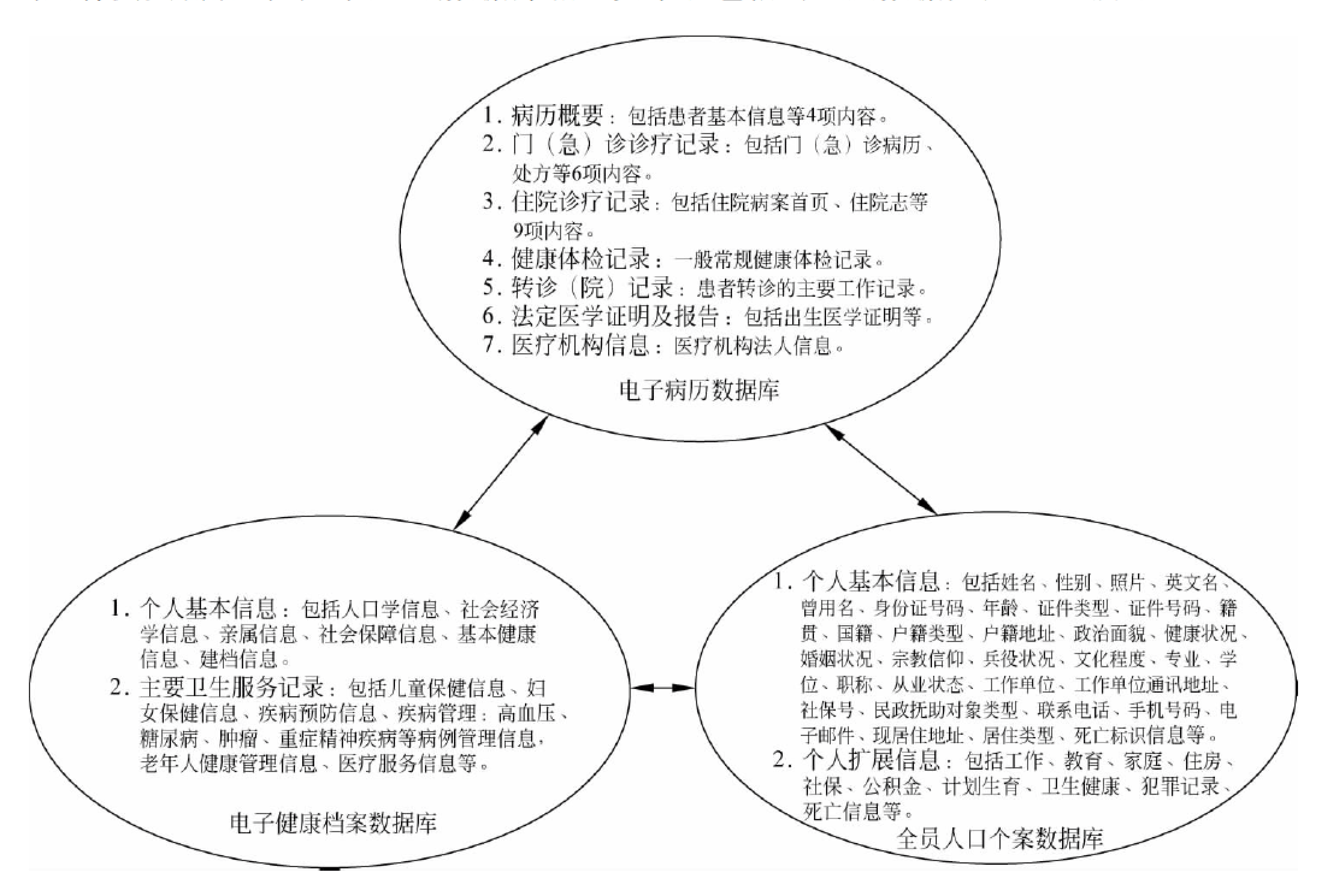


图8-14　医疗健康基础数据库图

电子健康档案的数据架构是以人的健康为中心，以生命阶段、健康和疾病问题、卫生服务活动（或干预措施）作为3个纬度构建的一个逻辑架构，用于全面、有效、多视角地描述健康档案的组成结构以及复杂信息间的内在联系。通过一定的时序性、层次性和逻辑性，将人一生中面临的健康和疾病问题、针对性的卫生服务活动（或干预措施）以及所记录的相关信息有机地关联起来，并对所记录的海量信息进行科学分类和抽象描述，使之系统化、条理化和结构化。

个人健康档案的三维概念模型，可以清晰地反映出每个个人不同生命阶段、主要疾病和健康问题、主要卫生服务活动三者之间的相互联系。同时，坐标轴上的三维坐标连线交叉所圈定的空间位置（域），表示了人在特定生命时期、因特定健康问题而发生的特定卫生服务活动所需记录的特定记录项集。由于三维空间中的任意一个空间位置都对应着某个特定的健康记录，从而构成了一个完整、立体的健康记录，这些健康记录全面地反映了个人健康档案内容的全貌，如图8-15所示。

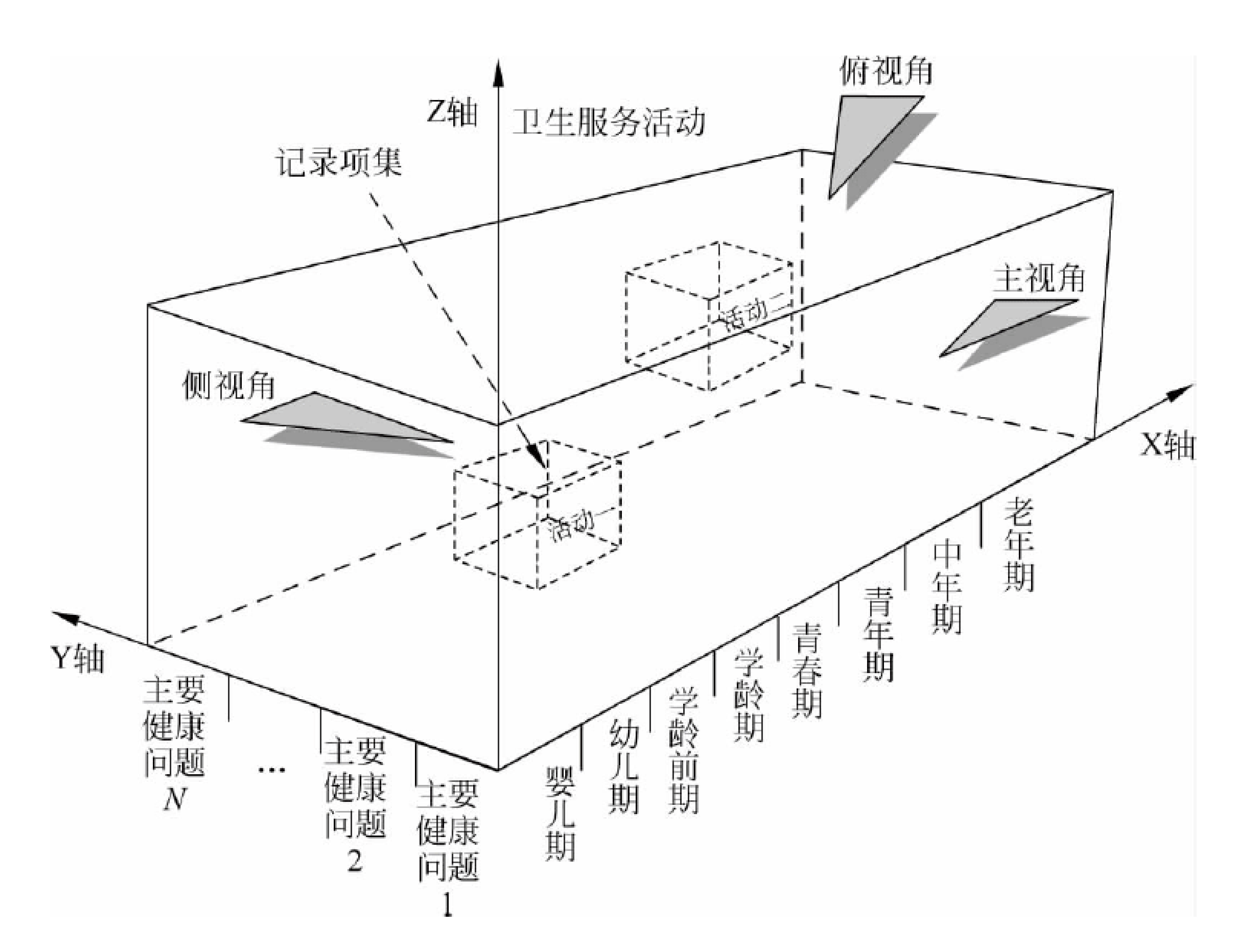


图 8-15 个人健康档案立体模型图

1. 第一维(X 轴):生命周期各个阶段

按照不同生理年龄可将人的整个生命进程划分为连续的若干生命阶段,例如:婴儿期(0～1 岁)、幼儿期(1～3 岁)、学龄前期(3～6 岁)、学龄期(6～12 岁)、青春期(12～20 岁)、青年期(21～45 岁)、中年期(46～60 岁)、老年期(60 岁以上)等 8 个生命阶段。也可以根据基层实际工作的需要,将人群划分为:儿童、青少年、育龄妇女、中年和老年人。

2. 第二维(Y 轴):健康和疾病问题

如图 8-16 所示,每一个人在不同生命阶段所面临的健康和疾病问题不尽相同。确定不同生命阶段的主要健康和疾病问题及其优先领域,是客观反映居民卫生服务需求、进行健康管理的重要环节。

3. 第三维(Z 轴):卫生服务活动(或干预措施)

针对特定的健康和疾病问题,医疗卫生机构开展一系列预防、医疗、保健、康复、健康教育等卫生服务活动(或干预措施),这些活动反映了居民健康需求的满足程度和卫生服务利用情况。

8.5.3 相关数据特征对比分析

从医药医疗健康大数据分析应用角度,本平台需要一个尽可能全和细的数据集合,所以理想状态是结合以上两部分数据内容形成的超集集合,甚至包括一些非医疗健康数据,如考察研究某种药对某种疾病的医疗效果时,如果能获得当地的气象天气信息,可能分析出的结果将明显不同。另外可以看出目前所给数据都是结构化数据,如果从大数据分析应用角度,理想的数据还应该包括图像、图形、文本等半结构和非结构数据,以及非关系数据(多维数据),才能构成满足医药医疗健康大数据分析应用的需求。

2000 年以来,我国的医疗数据的生成和采集主要局限于各大医院。近几年,随着社区

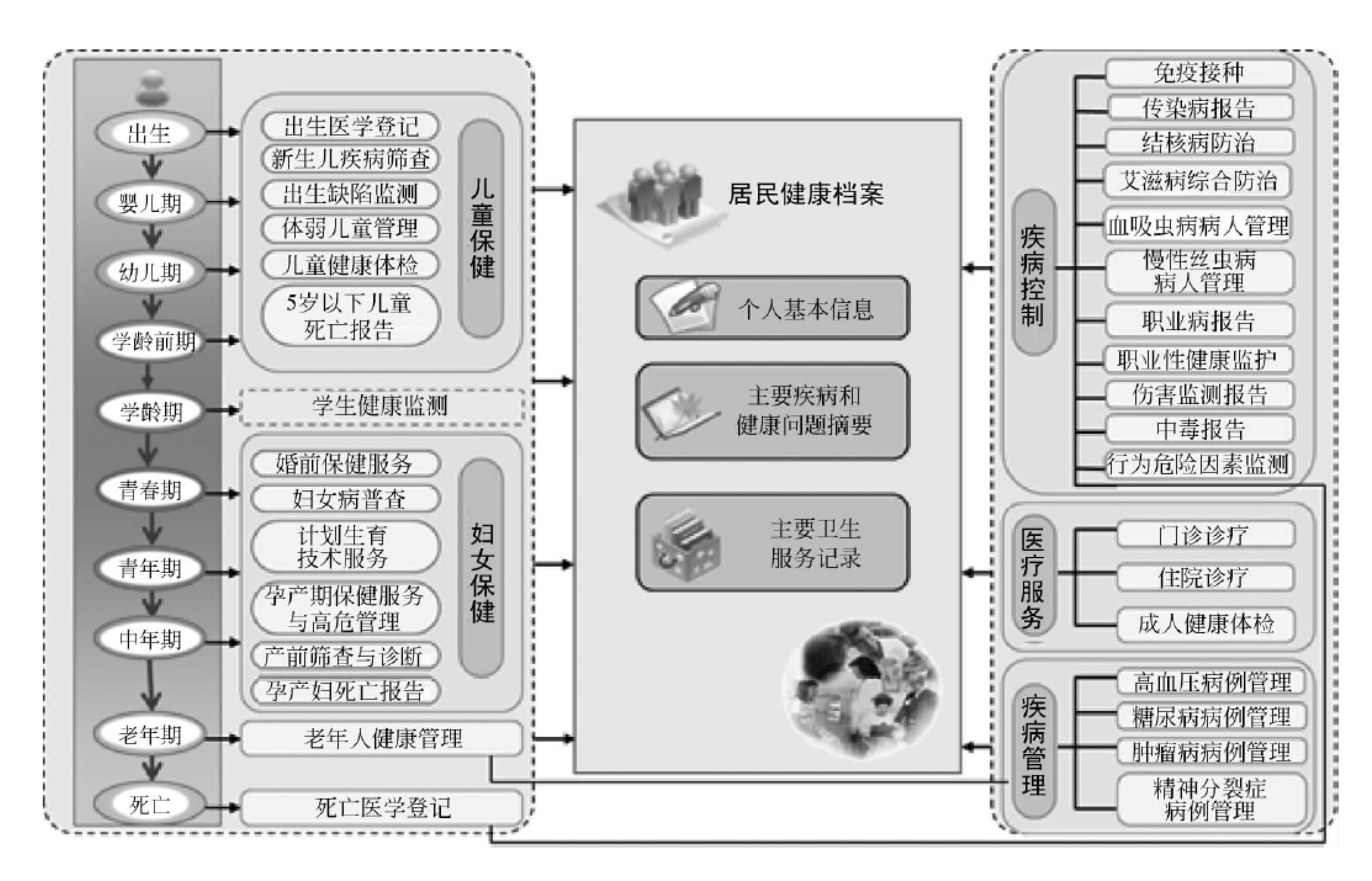


图 8-16　个人健康档案(EHR)健康与疾病管理

系统、新农合系统、村卫生室系统等基层医疗卫生信息系统逐步上线，医疗卫生数据源头也越来越多，数据量越来越大。从卫生服务的类型看，区域卫生信息的类型主要有：医疗服务类、公共卫生服务类、社区卫生服务类、卫生业务类、卫生管理服务类数据。根据估算，中国一个中等城市(一千万人口规模)50 年所积累的医疗卫生数据量就会达到 10PB 级。随着各地区域卫生信息平台的建设，存储于各医疗卫生机构的数据将逐步通过各种方式实现整合与共享。

由于医疗数据是多种数据源数据的汇总，数据之间的关系非常复杂，如图 8-17 所示。以患者为中心的服务需要把一个患者的全周期数据按照时间轴排列，并分析诊断、用药和患者生命体征、检验检测值之间的关联；以医生为中心的服务又需要把与医生相关的患者数据挑拣出来，并进行分类；以科室为中心的服务可能需要既从科室所属医生的角度，又要从在该科室就诊患者的角度进行分析；针对社区的服务可能需要统计整个社区居民某项指标(比如血压、血糖)的达标率。医疗数据的多维度、多粒度为各种信息服务的多角度、多层次分析提供了可能，但同时也为大数据分析带来了挑战。因为不可能为每一种信息服务存储一份特定的优化模式的数据，况且也无法枚举出所有可能的信息服务需求。这就需要医疗数据的存储模型能够适应灵活多变的多维统计分析需求。

8.5.4　临床信息学大数据分析

临床信息系统收集并处理各类数据来进行大数据分析实现医疗决策分析，包括诊断和治疗决策、科学研究和科学发现、临床试验、疗法趋势的监测、临床实践中不良反应的监测，还支持诸如对保险索赔和解的审计、欺诈识别等许多商业功能。数据来源主要来自于医院临床报告、工作会议记录、会诊报告、日记、临床信息系统数据、其他相关数据等，给临床医学

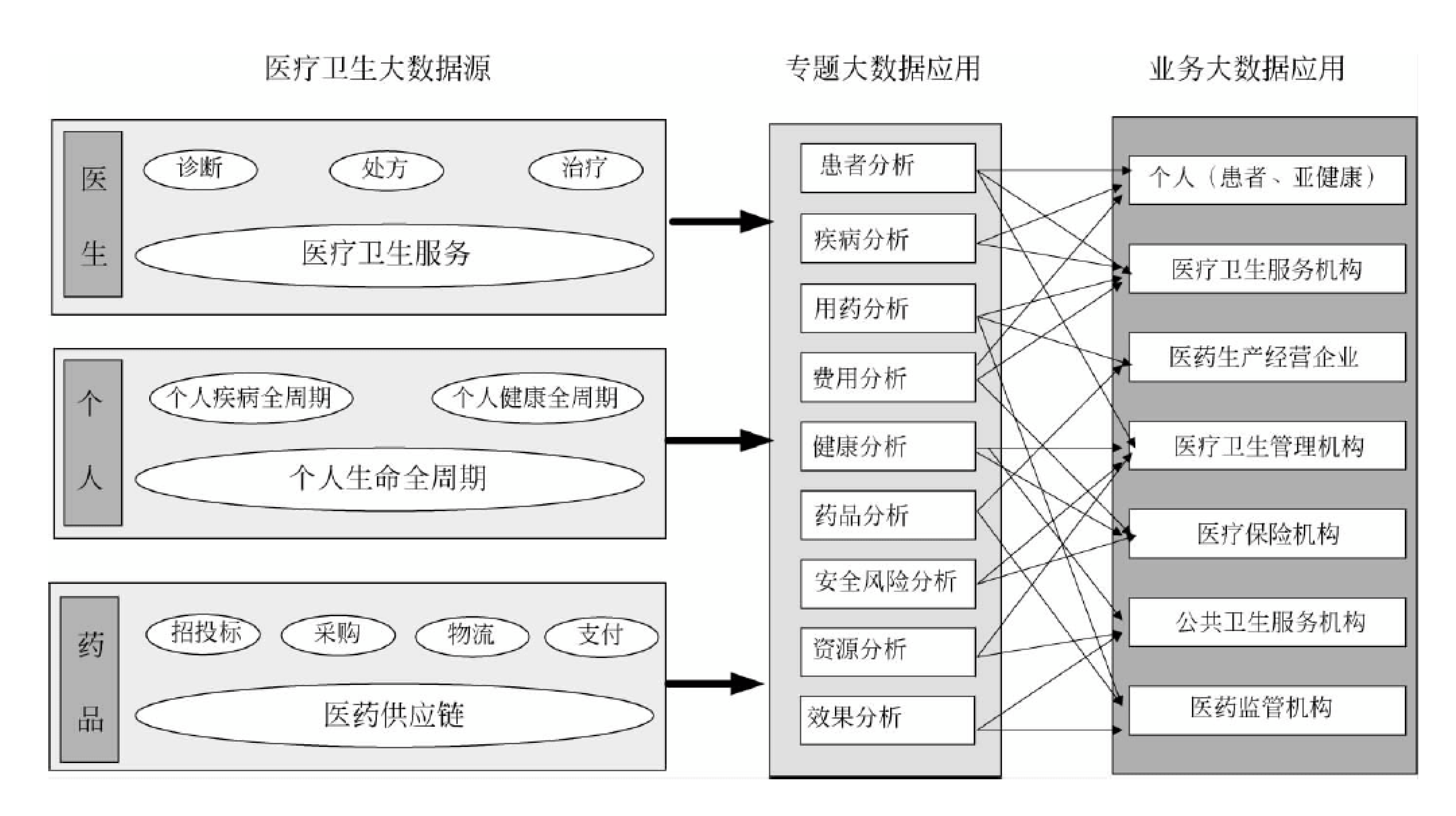


图 8-17　医疗多维数据分析图

大数据分析提供重要依据。

临床信息学所能实现的重要目标就是使物理学家和护士们具备监测工具功能，能够在病人诊治过程中提前发布病情概率预警和治疗方法提醒，从而医生能够在不良情况真正发生前采取预防措施。以下是一个真实的案例证实了数据采样分析的有效性。通过识别高频生命体征数据（包括心电图信号、血压、含氧量，以及相类似的以 O(125Hz) 频率测量的波形数据），这些数据主要采集于特级护理病人的床边。为了从生命体征数据中发现有意义的信息，每一条测量通道一开始都被分成 k 个连续观察的序列，接着使用 Fourier 变换以获取每一段原始信号的谱分布。然后，从那些医学上认为是良性观察序列中提取而得的多个谱分布，被聚集起来后形成一张 k 维常量表。之后对该表采用主成分分析，进一步考虑前 p 个成分，这 p 个成分构成了一个所给生命体征预期正常动态的零空间谱模型，给每条测量通道分别建立一个零空间模型。之后，对每一组新观察到的 k 个连续测量值进行 Fourier 变换，再投影到相应零空间模型的 p 个主成分上。随着对病人的观察，这些投影在每条测量通道中产生了 p 个时间序列。然后，对每个这样的时间序列应用控制图表的方法（使用累积和方法），并在 CuSum 警戒有所提高之处做时间标记。当所观测的生命体征的谱分解与基于普通医疗数据分布的期望不相吻合时，这些警戒就可以把这种情况标记下来。如果这样的标记超过 100 个，那就需要检查身体，可能会有健康恶化的情况发生。每一类自动提取的情况的预测效果是通过训练数据来量化的，这些数据除了包含生命体征外，还含有真实的健康预警作用。

8.5.5　医学文献研究知识发现

医学文献系统的文本挖掘是医学研究的重要课题之一。本节从神经文献出发结合神经本体原型中的条目概念和条目关系进行分析，同时文本挖掘得到的新知识又可以为神经本体更新和维护提供知识参考。文本挖掘可以用来对文献数据源进行自动搜索，搜索神经影像及脑功能相关方面的知识和信息，并挖掘潜在相关文献，从而有利于 NILS 的文献更新。

更为重要的是文本挖掘可以用来进行基于神经信息文献的知识发现，并且为将来假设驱动的研究模式提供基础平台，揭示文献中的多种隐含关系，为神经影像诊断和功能识别提供新的准确预测工具。随着学科专业的逐步细化，专业文献的研究范围开始逐渐缩小，专业间的沟通变得越来越困难。原本在专业文献间有价值的关联信息，由于专业文献的高度分化，日益被专业内部海量的信息掩盖，而不为研究人员所发现。由于公开发表的文献中存在着“未被发觉的公开知识”，重拾文献中的隐性关联，对于科学发现有着重大的意义。“基于文献的发现”就是以揭示蕴含于公开发表的文献，但尚未被人们认识或发觉的知识片段间的逻辑联系，从而提出知识假设，以便专业研究人员进一步证实，促使新知识的产生为目的的信息学研究。它是一个将表面上没有任何联系的文献中的具有隐含逻辑关系的知识片段组织起来的信息处理过程。虽然“基于文献的发现”不能取代传统的经验性科研工作以及文献查询工作，但它为科研人员提供了能够更容易地组织大量潜在有用信息的新手段，并且可以直接促使新知识的产生。我们设计的基于 Swanson 的知识发现方法，主要用于建立面向神经影像研究领域的交互式的医学知识发现支持系统。我们遵循的思路是首先找出与开始概念 A 相关的所有概念集 B(假设 A 是一种疾病，那么 B 可能是病理功能、症状等)，然后找出与 B 相关的所有概念集 C(如果 B 是病理功能，那么 C 可能是某个分子，从病理生理学方面与 B 有结构和功能上的关联)，最后检查 A 和 C 是否在医学文献中同时出现。如果 A 与 C 没有在同一文献中同时出现，我们就发现了二者之间新的潜在相关关系，并且这种关系将根据 A、C 各自特性通过实验方法、临床研究及人类知识判断加以验证或否决。与传统的知识发现方法相比，这一 ABC 模式的知识发现过程明显增强了目的性和方向性，它使科研人员找寻这种隐藏关系的过程不再盲目。B 的出现为科研人员提供有益的启发和关键性的引导，帮助专业研究人员认识和发现潜在有用的知识片段间的关联，进一步证实科学假设的可行性。ABC 的知识发现模式中 AB、BC 关联的找寻及 AC 关联的最终确证都需要强有力的技术和方法支持，这就需要用到文本挖掘。神经信息文献系统中的文本挖掘模块分为文本分类、命名实体识别、信息提取及知识发现 4 部分。

1. 文本分类

文本分类的方法可以借鉴于机器学习(Machine Learning)，常用的有简单贝叶斯分类法(Naive Bayes)、决策树(Decision Trees)、神经网络(Neural Networks)、最近似然法(Nearest Neighbor)、支持向量机(Support Vector Machines, SVM)等。在所有这些方法中都是用预先分类的文档集来进行训练，产生一个关于词或短语使用的统计模型，然后将此模型应用到未分类文档。在产生训练集及实际分类前有两个预备步骤：特征提取及特征集转换。文档描述可以基于字词(最常用)、词组合、字符顺序或与词发生频率联系的概念(很少使用)。特征集转换有两个目的：一是缩减特征集的大小，希望在改善效力的同时改善效率，二是对特征集进行缩放或增加权重来改进与所有文档集有关的文档特征描述。缩减特征集的大小通常采用词干法、排除禁用词及除去不能提高分类器的辨别力却增加负担的稀有词等方法。文本分类的评估指标主要有分类正确率和查全率、查准率。分类正确率是针对多分类系统的，而查全率和查准率主要是针对双分类系统的，可以对系统进行微调：牺牲查全率来提高查准率或牺牲查准率来提高查全率。神经信息文献的分类主要是将采集到的文献自动归档到神经系统结构、神经生理学、脑的整合功能、分子神经科学、临床神经科学等 8 大类及其相应小类中，既有利于 NILS 文献的自动更新，又为知识发现的下几步工作打下

基础。文本分类根据具体的事务需求可选择上述不同的分类方法,也可以将几种方法综合利用,直至完全满足需求。

2. 命名实体识别

生物医学文献中的命名实体,如基因、蛋白质、化合物及疾病等的识别,是促进相关文本的搜索及生物学实体间相互关系识别的关键。由于生物医学语言及词汇的复杂性及迅速发展,使得生物学实体的识别非常棘手。另外,由于这些术语及词汇本身缺乏统一的命名规范,必须通过上下文背景分析才能明白相同字符所表示的不同含义及同一术语不同的表达形式和别名等,这也增加了命名实体识别的难度。如 EGFR 既可指表皮生长因子(Epidermal Growth Factor Receptor),又可指估计肾小球过滤速度(Estimated Glomerular Filtration Rate),必须根据上下文来进行语义判断。神经信息文献系统的实体主要是指与神经影像及脑功能相关的术语,包括脑解剖结构实体(中枢神经系统、脑、小脑、浦肯野氏细胞)、神经系统疾病实体(脑膜瘤、癫痫、流行性脑炎)、脑的高级功能实体(学习、记忆、疼痛、嗅觉)、神经影像技术实体(正电子发射断层扫描术 PET、功能磁共振成像 fMRI)等不同层级的实体对象。识别命名实体识别模式的建立有手工方式和通过专家系统自动学习方式。机器学习技术、隐马尔可夫模型(Hidden Markov Models,HMMS)、贝叶斯学习、决策树、支持向量机、归纳法规则学习是命名实体识别中通常采用的方法。例如,HMMs 可以将基于词典的学习及上下文背景分析结合起来对实体进行标记。

3. 信息提取及知识发现

信息提取就是在自由文本中采用基于词类(Part-Of-Speech,POS)信息、本体或识别模式的方法识别出有生物学意义的实体关系和语义结构,如在分子生物学中识别出蛋白质相互作用。神经信息文献系统中的信息提取主要是提取神经影像、脑结构、脑功能、神经疾病、神经生理等之间的相互依存、互相关联及互为因果等多种关系。信息提取所得到的只是知识片断,还必须经过广泛、深入的分析和推导才能得出有用的综合知识信息,这些有用知识既可用于神经本体原型的构建又可用于大量非相关文献的知识发现。

8.6 医疗健康大数据展望

基于医疗卫生的海量数据,通过大数据分析进行预测具有非常广泛的市场应用前景,虽然现在说对医疗卫生产生颠覆式变革还为时尚早,但是基于医疗卫生信息的大数据分析将改变医疗卫生业务的方方面面并不为过。

未来医疗的精髓在于电子病历、电子健康卡以及相关信息(医药、人口等)的快速准确收集、传输、存储和分析处理,电子病历系统以电子化方式记录患者就诊的信息。

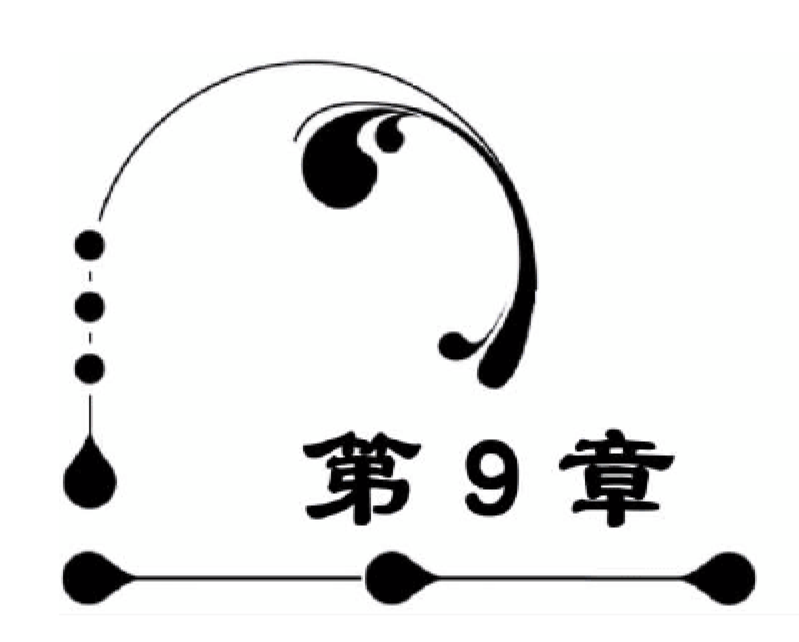
世界各国对电子病历的建设都极其重视,美国、日本、欧洲对电子病历的建设均进行了大量投入。2009 年美国通过的经济复兴法案同时包括 10 年 190 亿美元在电子病历领域的投入,目前的估计是实际投入将达 270 亿美元;英国政府 10 年投入了 55 亿英镑做电子病历。当数百万、千万的病历汇集在一起,利用大数据进行挖掘后,其应用前景十分惊人。

对患者来说,电子病历使患者拥有自己完整的电子健康和医疗档案,并可以通过索引在各个医疗机构调取自己的相关信息,实现跨地区、跨机构、终生的医疗健康信息共享。

对医疗机构来说，可以实现患者统一高效的管理。对于了解病情、临床决策、提高医疗质量及科学研究等都起到至关重要的作用。同时可以实现区域内不同医疗机构之间、不同应用系统之间的患者映射，确保患者信息交换的一致性和准确性。

对社保机构而言，可以通过患者主索引查阅患者的健康档案，从而准确地了解患者完整的医疗信息，为医疗保险提供确切的证明。

将电子病历信息进行大数据挖掘后，还会有更大的魔力。比如医疗信息系统会提醒医生开处方时患者的药物过敏反应。医疗信息系统还可用于人群监测，如对将会流行的传染病的早期症状加以监控，或对新上市的处方药的副作用加以关注。



第9章 环保行业大数据解决方案

环境领域将迎来一个大数据互联时代。若要全面呈现环境问题，尤其需要通过互联网实现环境数据、信息等要素互通共享，从而推动环境问题得到整体有效解决。具体来看，目前主要存在以下 3 种与环境相关的数据来源。

第一，环境质量。这是指外部自然环境质量表征，典型数据信息包括大气、地表水、水资源、土壤、辐射、声、气象等环境质量，通常由政府及有关部门（如环境保护部）公开其制作或获取的环境信息。

基于已经建立起来的以国控、省控、市控 3 级为主的环境质量监测网，形成信息公开机制，初步勾勒出了我国整体环境质量状况。比如，全国城市空气质量日报/时报（367 个城市）、全国主要流域重点断面水质自动监测周报（145 个监测断面）、全国辐射环境自动监测站空气吸收剂量率（44 个站点）等。

第二，污染源排放。这是造成环境污染的核心原因，具体体现为废水、废气、固废、放射源等形式，主要包括污染源基本情况、污染源监测、设施运行、总量控制、污染防治、排污费征收、监察执法、行政处罚、环境应急等环境监管信息。

《全国污染源普查公报》中的排污数据及信息，将是政府监管以及公众监督的重要前提与基础。目前，各地正逐步落实环境保护部出台的《关于加强污染源环境监管信息公开工作的通知》等文件。以北京市为例，虽然已按季度发布国控企业污染源监督性监测情况，而 27 家重点排污单位和上市企业仅于 2015 年起初步实现自行监测信息对外发布，实时信息公开仍无法实现。

第三，个人活动产生的与环境相关的数据信息，如用水量、用电量、生活中产生的废弃物等。尽管这些数据拥有巨大的潜在价值，但其分布却呈现天然的分散状态，互联网特别是移动互联网的快速广泛应用正在使上述信息的收集利用变得可行。

大数据的核心价值之一就是个性化的商业未来，是对人的终极关怀。环保电力大数据通过对市场个性化需求和企业自身良性发展的挖掘和满足，重塑中国电力工业核心价值，驱动电力企业从“以人为本”的高度重新审视自己的核心价值，由“以电力生产为中心”向“以客户为中心”转变，并将其最终落脚在“如何更好地服务于全社会”这一根本任务上。同时，电力大数据通过对电力系统生产运行方式的优化、对间歇式可再生能源的消纳以及对全社会节能减排观念的引导，能够推动中国电力工业由高耗能、高排放、低效率的粗放发展方式向低耗能、低排放、高效率的绿色发展方式转变。

9.1 环保物联网

环保物联网是物联网技术在环保领域的智能应用，通过综合应用传感器、全球定位系统、视频监控、卫星遥感、红外探测、射频识别等装置与技术，实时采集污染源、环境质量、生态等信息，构建全方位、多层次、全覆盖的生态环境监测网络，推动环境信息资源高效、精准地传递，通过构建海量数据资源中心和统一的服务支撑平台，支持污染源监控、环境质量监测、监督执法及管理决策等环保业务的全程智能，从而达到促进污染减排与环境风险防范、培育环保战略性新型产业、促进生态文明建设和环保事业科学发展的目的。

9.1.1 物联网概念

“物联网”这一概念在 20 世纪 90 年代就出现了，但是由于当时无线网络和传感器等相关技术还尚未成熟，因此没有引起普遍重视。2005 年，国际电信联盟在信息社会世界峰会上发布了《ITU 互联网报告 2005：物联网》报告，才正式引用了物联网概念。

物联网，即 Internet of Things(IoT)，顾名思义，就是“物与物相联构成的网络”。即通过射频识别、红外感应器、全球定位系统、激光扫描器等信息传感设备，按约定的协议，把任何物品与互联网相连接，进行信息交换和通信，以实现对物品的智能化识别、定位、跟踪、监控和管理的一种网络。得益于传感技术、网络通信技术、大数据、云服务等软件技术的发展，网络将从对计算机之间的互相连接，扩展到将每个实际物体连接起来。人与人、人与物、物与物之间能够互相交换信息。物体也可以灵活地参与到商业、信息和社会财产活动中。它们可以与环境进行互动，对环境的改变自动做出相应的响应。最终，将无缝地为人类的生产生活提供智能化和便捷化服务。

按照国际电信联盟(ITU)的定义：物联网是通过 RFID 和智能计算等技术实现全世界设备互联的网络。如图 9-1 所示，在不久的将来，物联网有可能如互联网一样，形成一个全球性的网络，在任何时间、任何地点，任何人和物都能建立连接。在互联网时代，主要强调的是任何时间、任何地点两个维度。在物联网时代，增加了第三个维度，强调了任何人和物体能够进行连接。

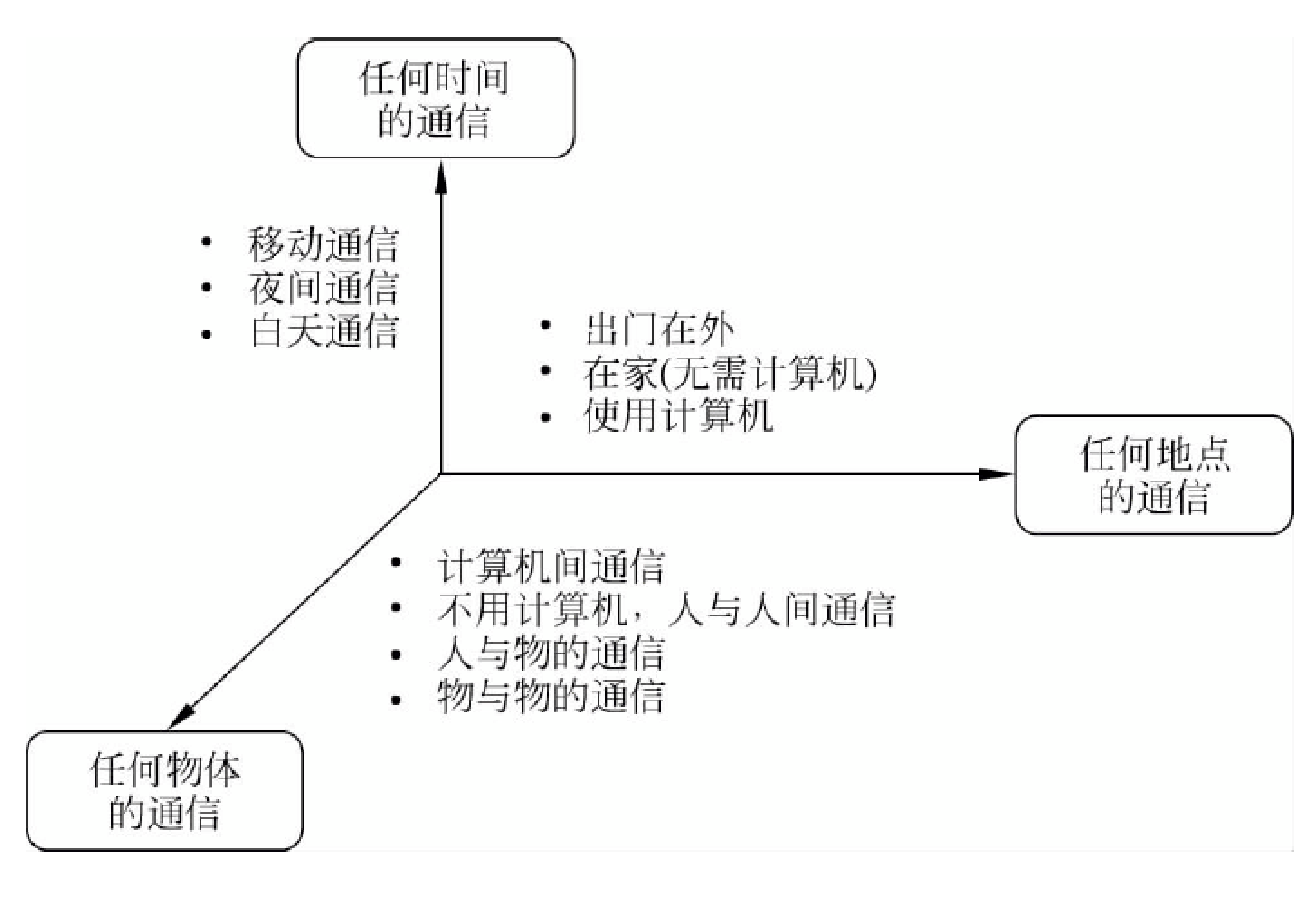


图 9-1 物联网新维度

9.1.2 物联网基本架构

物联网是新一代信息技术的高度集成和综合运用。它是基于社会、经济领域的实际管理和应用需求，利用感知技术和智能装置对物理世界进行感知识别，通过互联网、移动通信网等网络的传输互联，进行计算、处理和知识挖掘，实现人与物、物与物之间的信息交互和无缝连接，提升人对物理世界实时控制、精确管理和资源优化配置能力，从而实现生产生活的科学智能决策。在环保行业物联网中，主要使用到以下技术，下面分别具体介绍。

射频识别（Radio Frequency Identification，RFID）技术，又称无线射频识别，是一种通信技术，可通过无线电信号识别特定目标并读写相关数据，而无须识别系统与特定目标之间建立机械或光学接触。RFID 是一种能够让物品“开口说话”的技术，其标签中存储着规范而具有互用性的信息，通过有线或无线的方式把它们自动采集到中央信息系统，实现物品的识别，进而通过开放式的计算机网络实现信息交换和共享，实现对物品的“透明”管理。RFID 系统主要由 3 部分组成：电子标签（Tag）、读写器（Reader）和天线（Antenna）。其中，电子标签芯片具有数据存储区，用于存储待识别物品的标识信息；读写器是将约定格式的待识别物品的标识信息写入电子标签的存储区中（写入功能），或在读写器的阅读范围内以无接触的方式将电子标签内保存的信息读取出来（读出功能）；天线用于发射和接收射频信号，往往内置在电子标签和读写器中。RFID 具有无须接触、自动化程度高、耐用可靠、识别速度快、适应各种工作环境、可实现高速和多标签同时识别等优势，因此可用于广泛的领域。

条形码/二维码是用某种特定的集合图形按一定规律在平面分布黑白相间的图形记录数据符号信息的。其在代码编制上巧妙地利用构成计算机内部逻辑基础的“0”“1”比特流的概念，使用若干个与二进制相对应的几何形体来表示文字数值信息，通过图像输入设备或光电扫描设备自动识读以实现信息自动处理。同时还具有对不同行的信息自动识别功能，及处理图形旋转变化等特点。目前，条形码/二维码技术已经有了相当广泛的应用。

传感器是一种检测装置，能感受到被测量的信息，并能将感受到的信息，按一定规律变换成为电信号或其他所需形式的信息输出，以满足信息的传输、处理、存储、显示、记录和控制等要求。它是实现自动检测和自动控制的首要环节。传感器的存在和发展，让物体有了触觉、味觉和嗅觉等感官，让物体慢慢变得活了起来。通常根据其基本感知功能分为热敏元件、光敏元件、气敏元件、力敏元件、磁敏元件、湿敏元件、声敏元件、放射线敏感元件、色敏元件和味敏元件等 10 大类。

摄像头一般具有视频摄像/传播和静态图像捕捉等基本功能，它是借由镜头采集图像后，由摄像头内的感光组件电路及控制组件对图像进行处理并转换成计算机所能识别的数字信号，然后借由并行端口或 USB 连接输入到计算机后由软件再进行图像还原。目前，摄像装置被广泛用于各类监控系统中。

环保网络包括有线/无线通信网、互联网、物联网等。

有线/无线通信网主要用于企业内部的数据交换以及各感知设备收集数据的回传处理等，其主要依靠网络基础设施实现。在企业运行中，特别是保安行业这种比较机密、敏感的行业，大量的数据只能在企业内部流转，而不能部署在互联网上，而且不同的人员应当具有不同的接入权限，这些都需要内部通信网络来进行控制。因此企业内部的有线/无线通信网承担着企业内部信息和机密信息的流转工作，可以说相当重要，而且其对安全性的要求也很

高。此外，企业内部的资源、人员的调度和沟通也离不开有线/无线通信网的支持。

互联网大家都很熟悉，它是网络与网络之间所串联成的庞大网络，这些网络以一组通用的协议相连，形成逻辑上的单一巨大国际网络，主要依靠网络基础设施实现。互联网主要用于用户对企业提供的各类服务的访问以及企业获取外部信息等，在用户获取服务时，应同时提供网页端和移动端的互联网服务，以提升用户的使用体验，无论是用户发起某种需求或者请求其他服务，都应在网页端和移动端提供互联网接口，使用户方便地获取相应的服务。

物联网是一种利用局部网络或互联网等通信技术把传感器、控制器、机器、人员和物等通过新的方式联在一起，形成人与物、物与物相联，实现信息化、远程管理控制和智能化的网络。物联网是互联网的延伸，它包括互联网及互联网上所有的资源，兼容互联网所有的应用，但物联网中所有的元素（所有的设备、资源及通信等）都是个性化和私有化的。物联网主要依靠网络基础设施和各物联网基础设施协调实现。通过各感知器之间的信息通信和信息回传，企业可以获知系统内全部资源的运行状况、实时画面等，从而实现对系统内各资源的精细化、智能化管控。

如图 9-2 所示为环保监控物联网架构图。物联网建设内容包括感知层、网络层、应用层的建设。感知层包括条码识读器、RFID 读写器、传感器、摄像头、GPS、手机、实验室、智能车、条形码/二维码扫描器等感知设备的部署、接入和管理；感知设备通过传感网关、M2M 终端、互联网关将采集的信号发送给物联网网络层，物联网网络层建设包括物联网信息中心和物联网管理中心的建设，物联网信息中心是数据采集的信息库和计算能力集合，属于大数据平台；物联网管理中心主要针对物联网数据的统一编码、认证、鉴权和计费。通过信息中心和管理中心的基础物联网数据能力，实现物联网应用层的建设。

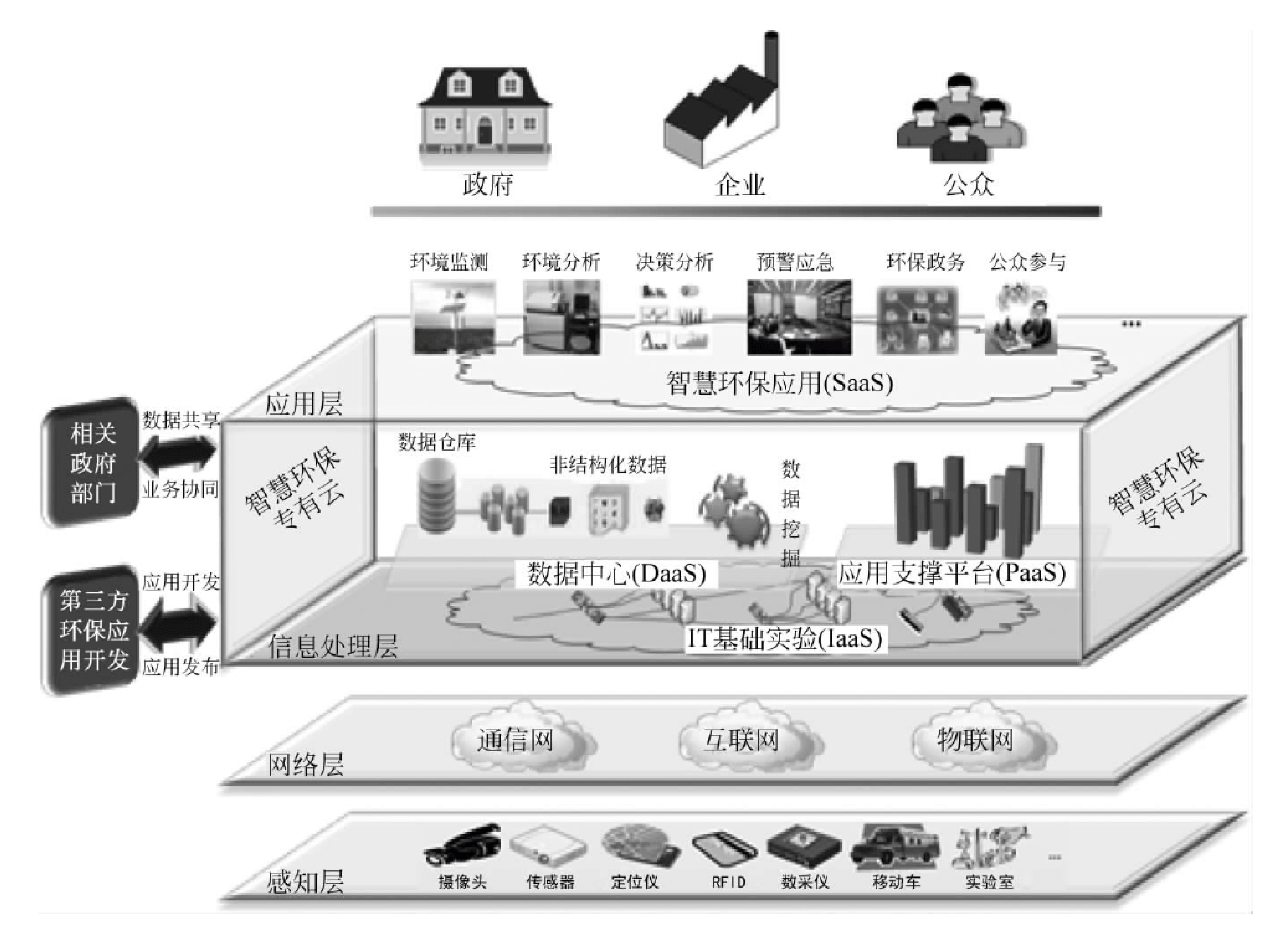


图 9-2　环保监控物联网架构图

9.1.3 环保物联网数据

在物联网大数据时代,互联网、移动互联网、物联网等产生的数据增长比以往任何一个时期都快很多,具有数据规模大、产生速度快、数据结构复杂多样这 3 个显著特性。随着物联网的快速发展,物联网中的数据的增长也非常迅猛,其创造出的数据将远多于互联网。物联网所创造出的数据,描绘的是物质运动、经济变化、自然变化等规律,其数据更加真实、可靠、有价值,可以从中挖掘出更丰富、更有用的知识。

9.2 环保电力脱硫

9.2.1 火电脱硫的重要性

中国环境科学研究院的研究表明,为使我国的工业可持续发展,从长远来看,我国二氧化硫的排放量应控制在 1200 万吨/年,其中电力行业排放的二氧化硫应控制在 550 万吨/年以下,新修订的《火电厂大气污染物排放标准》(GB 13223—2003)规定了火电厂大气污染物最高允许排放限值,火电厂建设项目的环境影响评价、设计、竣工验收和建成运行后的排放管理必须遵守本标准,因此,火电厂必须配备脱硫系统。

根据规定,2004 年后建成的火电站的二氧化硫的浓度必须低于 400mg/m^3。在脱硫效率方面,一般规定需要在 95%或以上。所以,为了使排放废气达到标准,必须对废气进行处理,而大部分火电站采用的就是废气脱硫系统(FGD)。

9.2.2 火电脱硫系统工作原理

脱硫系统属于电力系统的环保系统,主要是去除火电厂废气中的二氧化硫等硫化物,使其含硫量符合国家规定标准。

脱硫系统工作原理图,如图 9-3 所示。脱硫系统主要依靠石灰石与二氧化硫发生化学

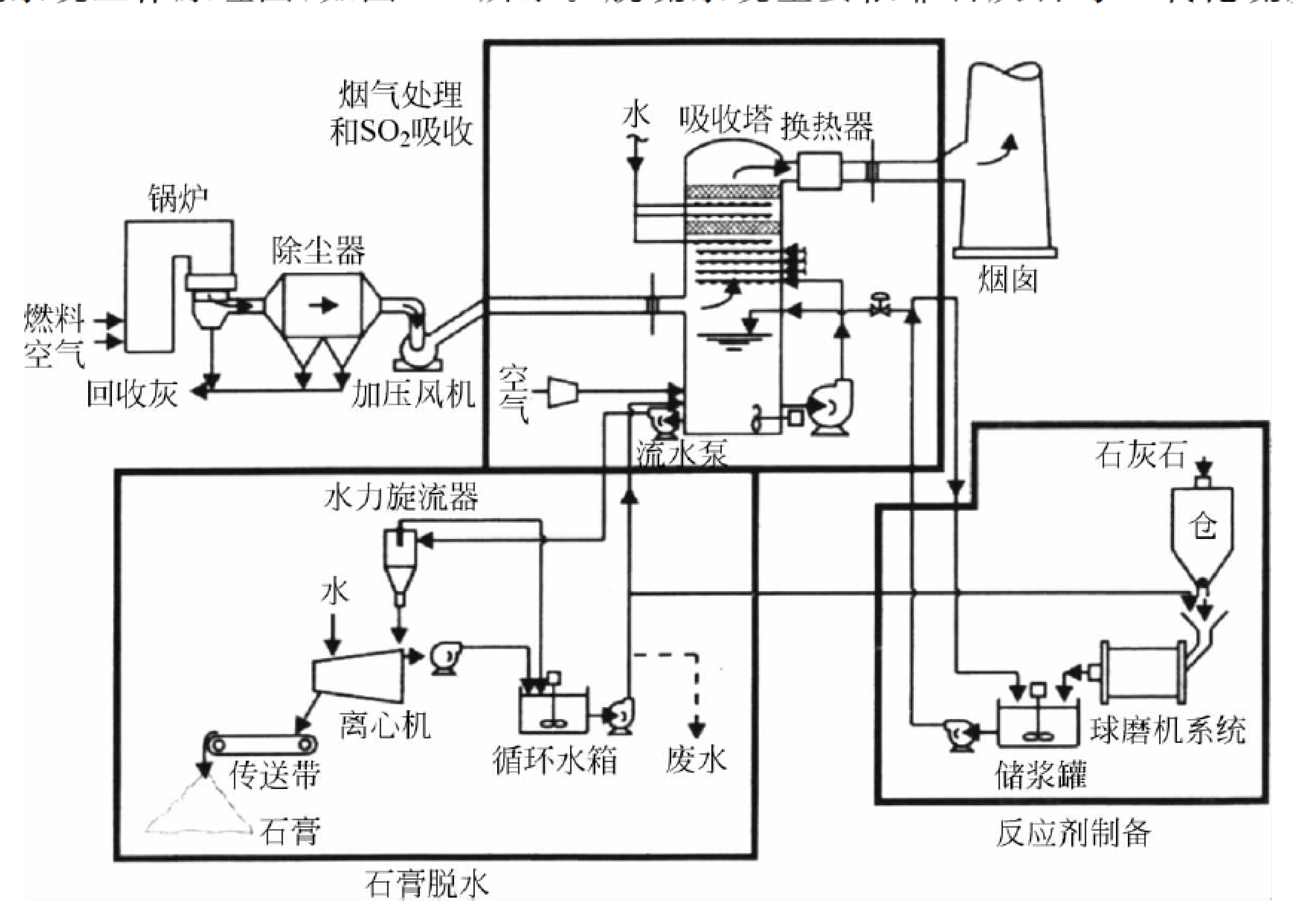


图 9-3 脱硫系统工作原理图

反应来达到脱硫的作用，该反应发生在吸收塔内，送入吸收塔的吸收剂——石灰石浆液与冷却后进入吸收塔内的烟气充分混合后，烟气中的二氧化硫(SO_2)与石灰石($CaCO_3$)及鼓入吸收塔内的氧气发生化学反应，生成二水硫酸钙，即石膏($CaSO_4 . H_2O$)，脱硫后的烟气依次经过除雾器除去雾滴，加热升温后经烟囱排入大气，而石膏则经过石膏浆泵排出吸收塔。

9.2.3 火电脱硫相关数据

火电厂大数据有非常重要的参考价值，我们希望通过数据分析，挖掘其中有价值的信息。通过火电厂数据分类，有锅炉数据(锅炉效率、过热气温、再热气温、排雾率、锅炉含氧量等)、汽机数据(汽机效率、汽耗率、真空度等)。

根据《电厂排口数据分析报告》和脱硫量相关的烟气系统参数分析，可知：旁路挡板开度，浆液 pH 值，增风压机电流，循环泵电流等。其他还有：进入吸收塔内烟气的温度、湿度、含氧量，机组负荷，吸收塔排出烟气的温度、湿度、含氧量等，如表 9-1 所示。

表 9-1 脱硫系统参数

烟气系统	公用系统	吸收塔系统	烟气换热器	氧化空气	增压风机
氧量	除雾器冲洗水流量	石膏浆液 pH 值	主驱动电机电流	一级轴承温度	轴承温度
烟温	石灰石浆液流量	吸收塔液位	顶部轴承油温	二级轴承温度	电机 A 相温度
含湿量	粉仓除粉器差压信号	循环泵轴承温度	转换信号	电机绕组温度	电机 B 相温度
炉旁路挡板开度	真空泵回路电流	循环泵电机线圈温度	底部轴承油温	驱动轴承温度	电机 C 相温度
机组负荷	工艺水泵电流	搅拌器电流	吹扫蒸汽温度	空气出口温度	轴承温度
泵液 pH 值	除雾器冲洗水泵电流	浆循环泵电流	吹扫蒸汽压力	空气出口流量	润滑油温度
增压风机电流	除雾器冲洗水压力	石膏排出泵电流	低泄漏风机电流	电流	润滑油箱温度
循环泵电流	循环排污水流量	石膏浆液密度	辅驱动电机电流	冷却水回水温度	亚油箱油温

由火电厂运营情况可知，进口含氧量基本处于 4%左右，出口会增加 0.5%。数据采集要在一定状态下进行，假设风机正常运行状态，不需要调节；炉旁路挡板开度基本保持关闭；浆液循环泵保持不能调；pH 值为监控到的状态参数，目前运行时一般调节在 5.1～5.5。因此，我们只计算增压风机和 pH 值。具体参考图 9-4，脱硫系统工艺流程图。

9.2.4 脱硫性能优化目标

通过对海量的脱硫历史数据进行分析，在满足国家规定的脱硫指标的前提下，对脱硫系统可调参数进行优化，实现降低脱硫成本的目标。

1. 脱硫参数优化

脱硫参数优化目标：在满足一定脱硫量和脱硫效率(95%)的前提下，对脱硫系统的各个可调参数进行优化。主要针对锅炉效率优化、排口脱硫量预测等。

2. 脱硫成本优化

脱硫成本优化目标：在满足一定脱硫量和脱硫效率(95%)的前提下，最小化脱硫系统

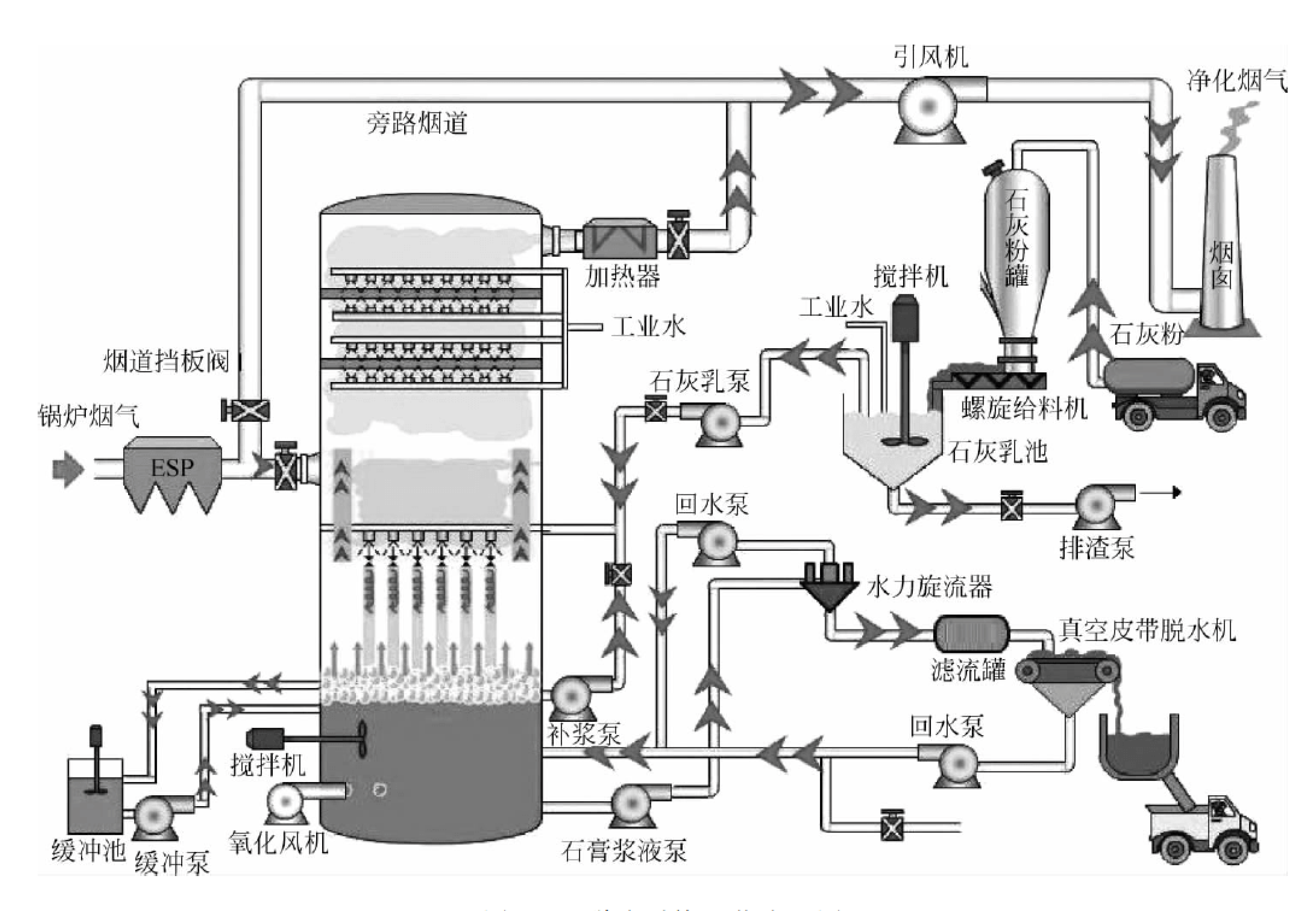


图 9-4　脱硫系统工艺流程图

成本，从而最大化经济利益。

9.3　火电行业脱硫大数据分析

随着信息通信技术的不断进步，数字化、信息化已经渗透进人们生活中的各个角落。据IDC编制的年度数字宇宙研究报告《从混沌中提取价值》表明，世界已进入了“数字摩尔时期”，全球数据量大约每两年翻一番。从人类出现文明到2003年，人类总共产生了5EB（百亿亿字节）数据，而这仅是当前人类社会两天的数据量。我们正处于数据世界一个重要历史爆发期的边缘，数据是资产是财富的观念已深入人心，大数据应用已是大势所趋，“大数据时代”已然到来。

9.3.1　主要理论和方法

随着大数据分析在各个行业应用的深入发展，基于大数据分析和数据挖掘的知识和方法主要包括：知识发现、机器学习、统计分析、模式识别和人工智能等领域的方法，如聚类、分类、关联规则分析、神经网络、遗传算法、进化算法和粗糙集等。以下主要介绍电力行业应用的大数据分析方法。

1. 聚类和模糊聚类

聚类就是将数据对象分组成为多个类或簇，同一个类中的对象具有较高的相似度，而不同类中的对象差别较大。一般情况下，聚类分析不要求训练数据提供类标记，聚类可以用于产生这种类标记。聚类按照某个特定标准最终形成的每个类，在空间上都是一

个稠密的区域，聚类技术可以把数据划分为一系列有意义的子集，进而实现对数据的分析。

（1）k 均值聚类。k 均值（k-means）算法是一种常用的基于划分的聚类算法。k 均值算法是以 k 为参数，把 n 个对象分成 k 个簇，使簇内具有较高相识度，而簇间相识度较低。k 均值算法的处理过程为：首先随机选择 k 个对象作为初始的 k 个簇的质心，然后将其余对象根据其余各个簇的质心的距离分配到最近的簇，最后重新计算各个簇的质心。不断重复此过程，直到目标函数最小为止。簇的质心为簇内所有点的算术平均值，对象到质心的距离一般采用欧式距离，目标函数采用平方误差准则函数

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_j} | P_j - m_i^2 |$$

其中，E 为数据库中所有对象与相应簇的质心的距离之和，P 代表对象空间中的一个对象，m 为簇的算术平均值。公式所示的目标函数旨在使所获得的聚类具有这样的特点：各聚类本身尽可能紧凑，而不是聚类之间尽可能地分离。

k 均值算法尝试找出使平方误差函数值最小的 k 个划分，是解决聚类问题的一种经典算法。它的主要优点是算法简捷，如果数据分布较均匀，结果簇是密集的，且簇与簇之间区别明显时，它的效果最好。在处理大的数据集时，该算法是相对可伸缩和高效率的。

（2）模糊聚类。模糊聚类算法包括 FCM 算法、模糊 C 均值算法等。FCM 算法是一种基于划分的聚类算法，它的思想就是使得被划分到同一簇的对象之间相识度最大，而不同簇之间的相识度最小。模糊 C 均值算法是普通 C 均值算法的优化，普通 C 均值算法对于数据的划分是硬性的，而 FCM 则是一种柔性的模糊划分。

2. 模糊关联规则挖掘

关联规则挖掘是从给定的数据集中发现频繁出现的项集模式知识，即从大量的数据中挖掘出有价值的描述数据项之间相互联系的有关知识。关联规则按处理的变量类别可分为：布尔型和数据型关联规则；按挖掘的抽象层次可分为：单层和多层关联规则；按用于挖掘的数据维度可分为：单维和多维关联规则。

9.3.2 最优化脱硫可调参数

通过数据挖掘分析对脱硫系统可调参数进行最优化处理，其过程主要分为 3 个阶段，如图 9-5 所示，包括 k-means 自然工况划分、FCM 模糊化聚类、模糊关联规则挖掘。



图 9-5 数据挖掘分析过程示意图

第一步，基于 k-means 的自然工况划分。

机组的负荷工况变化范围很大。负荷作为机组最重要的边界条件，它的变化会引起机组很多运行参数的变化。传统的优化方法往往是选择机组的几个典型负荷点作为典型工况进行研究，如 50%、70%、80%、90%等的负荷工况。目前，电厂经常面对诸如环保约束变化等情况，典型负荷不一定是其常见的运行工况。因此，本文提出从机组历史数据中分析该机

组常见的运行工况,采用聚类算法将负荷和入口 SO_2 含量进行自然划分。

第二步,基于 FCM 模糊化聚类的连续值离散化。

经过划分工况簇后,在每一个工况簇下求最优可调参数。

因为输入属性全是连续值,不能直接采用关联规则挖掘进行输入属性的最优化,所以需要对输入属性进行离散化处理。最简单的离散化方式是分段,但是这样做会造成硬边界,为了避免硬边界,我们采用基于模糊集的模糊 C 均值算法。该算法对于每一个输入属性,将其划分成 C 个类,每一个连续值对于每一个类都有一个隶属度,且该连续值对于 C 个类的属性度的和为 1。利用 FCM 算法就是求得输入数据中每一条记录在每一个输入属性的值对于每一个该输入属性的分类的隶属度。如果有 N 条输入记录,M 个输入属性,每个输入属性划分成 C 个类,则利用 FCM 算法求得的是一个 $N\times M\times C$ 的矩阵。最后将每一个属性模糊化为 3 个区间段概念,即高、中、低。

第三步,通过模糊关联规则挖掘最优参数组合。

连续值离散后,我们采取模糊关联规则挖掘的算法,计算每个工况簇下最优参数组合,即每个参数应该调为高、中或者低。最后,取聚类质心和区间代表该工况簇下该参数应该调节的大小。

9.3.3 最小化脱硫系统成本

要想达到最小化脱硫成本,首先要定义整个脱硫系统的成本(cost),cost 为各个可调参数的函数

$$\text{cost} = f(x)$$

其中,$x=\{x_1,x_2,\cdots,x_m\}$是各个可调参数。

定义脱硫量与各个可调参数的关系

$$\text{Gd} = g(x)$$

要使得脱硫量大于一个给定值 h,其中,h 为给定的最小脱硫量。

$$\text{Gd} = g(x) \geqslant h$$

因此,最小脱硫量可以表达为

$$\text{Min cost} = f(x)$$

使得

$$\text{Gd} = g(x) \geqslant h$$

可以先用模糊关联规则求得离散属性值的集合,然后对于所有隶属于这些离散属性值集合的运行记录求得其成本(cost),取成本最低的运行记录的值为最终结果。

在成本最低参数选取阶段,对于所有隶属于模糊关联规则挖掘出的离散属性集合的输入记录,将其连续值属性代入 cost 方程中,求出使得成本最低的记录。该条记录的各个连续值属性的值就是使得成本最低的参数。

9.4 空气质量大数据分析评价体系

一直以来,北京十分关注环境质量改善,尤其是 2008 年环境质量跨越式提升。但随着人口数量的攀升、城市规模的扩张,环境质量依然不能令人满意。2013 年 1 月,北京的

PM2.5 基本在 100 微克/立方米以上，多数时间在 139 微克/立方米以上，重度污染超过 4 次，仅 5 天没有雾霾。2012 年北京因为 PM2.5 早死的人数达到 2589 人，经济损失达 20.6 亿。作为发展中国家建设世界城市的先锋，客观上要求北京必须瞄准国际城市的高端形态，不断提高城市国际化水平，然而环境污染已经成为北京吸引跨国公司、国际组织和全球高端人才的一道鸿沟，严重阻碍了世界城市的建设步伐。北京环境污染不仅带来巨大的经济损失，也给市民的身体健康带来巨大威胁，已经引起了中央和北京市的高度重视。北京环境污染治理需要借鉴发达国家的先进经验，但更需要探索适合国情、市情的治理对策。因此，需要准确评价和正确认识北京在全国范围内的环境污染水平和北京自身的环境污染变化趋势，全面掌握北京环境污染现状，准确把握其区域差异和演变规律，这对于科学提出北京环境污染治理的政策建议，促进北京经济、社会可持续发展具有重要的现实意义，同时也对我国其他城市的环境污染治理政策的制定提供有益借鉴。

9.4.1 基于熵权的模糊综合评价方法的原理

模糊综合评价是以模糊数学为基础，对多种影响因素的事物或现象进行总的评价，克服了各种复杂多变的不确定性因素的影响。本文采用熵权法为评价指标客观赋权，熵权法是把评价中各评价指标的信息进行量化与综合，计算各指标反映的信息熵，通过各指标的信息熵来确定权重的客观赋权方法，熵权法有效地避免了人为因素的干扰，使评价结果更符合实际，从而给出客观、可靠的评价结果。设有 m 个评价指标，n 个评价对象，则形成原始数据矩阵 $X=(x_{ij})_{m\times n}$，对于某项指标 i，指标值 x_{ij} 的差异越大，则该指标在综合评价中所起的作用越大。如果某项指标的指标值全部相等，则该指标在综合评价中几乎不起作用。基于熵权法的模糊综合评价有以下 4 个步骤。

1. 原始数据矩阵标准化

m 个评价指标，n 个评价对象得到的原始数据矩阵为

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{9-1}$$

对该矩阵标准化得到

$$R=(r_{ij})_{m\times n} \tag{9-2}$$

式中，r_{ij} 为第 j 个评价对象在第 i 个评价指标上的标准值，$r_{ij}\in[0,1]$。其中对大者为优的收益性指标而言，有

$$r_{ij}=\frac{x_{ij}-\min\limits_{j}\{x_{ij}\}}{\max\limits_{j}\{x_{ij}\}-\min\limits_{j}\{x_{ij}\}} \tag{9-3}$$

其中对小者为优的成本性指标而言，有

$$r_{ij}=\frac{\max\limits_{j}\{x_{ij}\}-x_{ij}}{\max\limits_{j}\{x_{ij}\}-\min\limits_{j}\{x_{ij}\}} \tag{9-4}$$

2. 定义熵

环境质量统计口径包括 5 个方面：大气环境、水环境、声环境、辐射环境和生态环境。其中，31 个城市的辐射环境质量均属正常及以上；生态环境质量比本文研究的环境质量含

义更广，将另立专题展开专门研究。因此，本文将不涉及辐射环境和生态环境。如无特殊声明，本节所涉及的与北京相关的数据均通过《北京统计年鉴 2001—2012》和《2000－2011 年北京市环境状况公报》和《北京区域统计年鉴 2012》直接或计算整理获得。

其中，$f_{ij}=\frac{r_{ij}}{\sum_{j=1}^{n} r_{ij}}$，$k=\frac{1}{\ln n}$，当 $f_{ij}=0$ 时，令 $f_{ij}\ln f_{ij}=0$。

在有 m 个指标，n 个被评价对象的评估问题中，第 i 个指标的熵定义为

$$H_i=-k\sum_{j=1}^{n} f_{ij}\ln f_{ij},\quad i=1,2,\cdots,m \tag{9-5}$$

3. 定义熵权

定义了第 i 个指标的熵之后，可得到第 i 个指标的熵权定义，即

$$\omega_i=\frac{1-H_i}{m-\sum_{i=1}^{m} H_i} \tag{9-6}$$

其中，$0\leqslant\omega_i\leqslant 1$，$\sum_{i=1}^{m}\omega_i=1$。

4. 计算模糊综合评价结果

对于权重 $A=\{\omega_1,\omega_2,\cdots,\omega_n\}$ 与模糊关系矩阵 R，计算模糊评价结果

$$Z=A\circ R=[z_1,z_2,\cdots,z_m] \tag{9-7}$$

式中，“$\circ$”为模糊算子，Z 为综合评价值。

9.4.2 综合评价指标选择与数据来源

根据北京环境保护监测中心监测项目、《北京统计年鉴 2000—2013》和《2000—2012 北京环境状况公报》统计口径，结合我国和北京环境污染的现实情况，综合考虑大气环境、水环境和声环境 3 个维度，选取可吸入颗粒物每立方米含量、二氧化硫每立方米含量、二氧化氮每立方米含量、空气质量低于二级天数、化学需氧量（横向比较时为人均化学需氧量）、区域环境噪声和道路交通噪声等 7 个指标，以期能够全面、客观地反映环境污染水平。

在北京自身纵向比较评价中，原始数据为北京 2000—2012 年度的相关数据。7 个评价指标数据根据《北京统计年鉴 2001—2013》和《2000—2012 年北京市环境状况公报》计算整理所得。

在北京与其他城市横向比较评价中，原始数据为中国大陆 4 个直辖市和 27 个省会城市 2012 年度的相关数据。7 个评价指标数据根据《中国统计年鉴 2013》和 31 个城市的《2012 年国民经济和社会发展统计公报》计算整理所得。

9.4.3 环境质量综合评价结果及分析

各指标均属逆向指标，采用式(9-4)对全部指标数据进行标准化处理，利用基于熵权的模糊综合评价法对北京环境质量分别进行纵向和横向综合评价。计算结果如图 9-6 和表 9-2 所示。

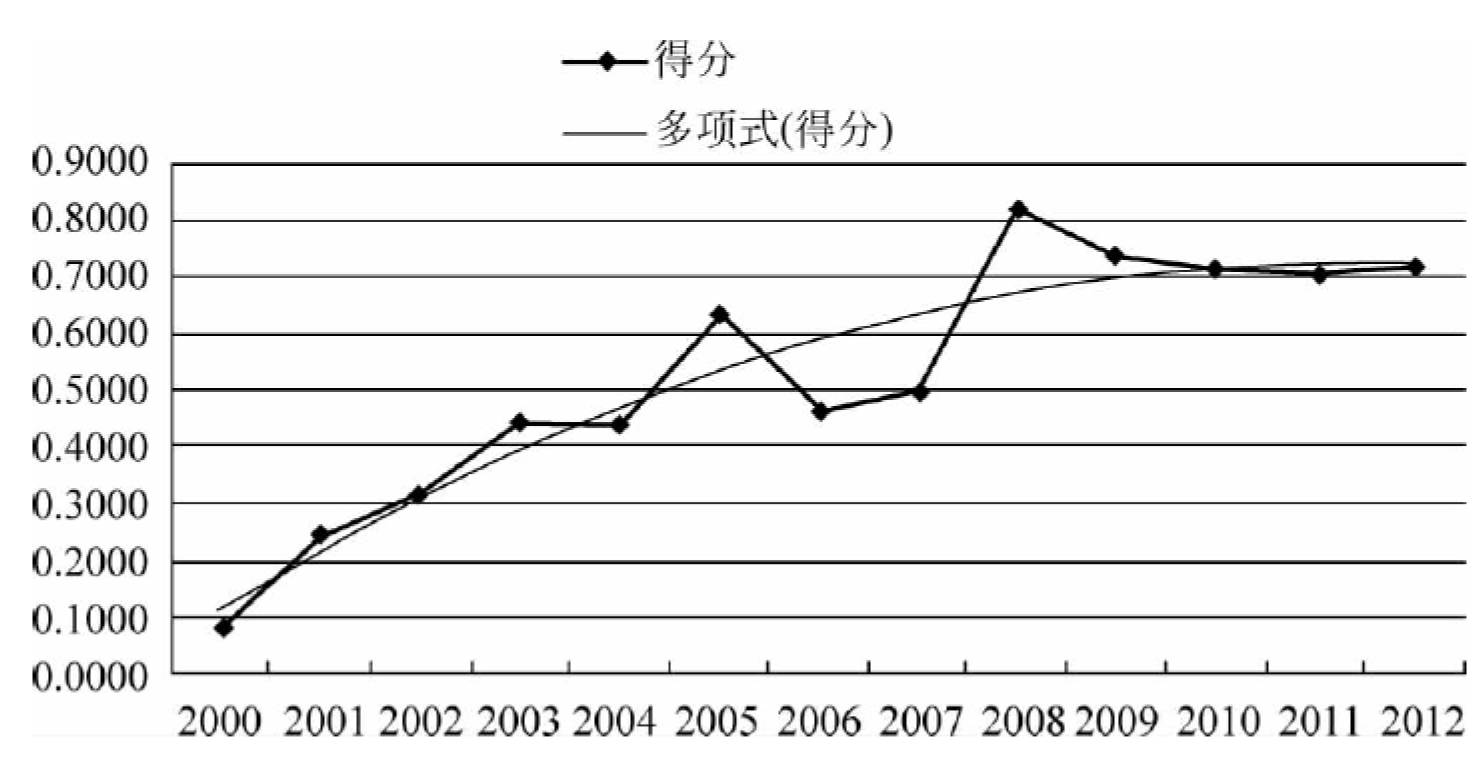


图 9-6　2000—2012 年北京环境质量变化趋势

表 9-2　全国 31 个城市环境质量综合评价结果排名

排名	城市	得分	排名	城市	得分	排名	城市	得分
1	海口	0.8310	12	上海	0.5640	23	西安	0.4401
2	拉萨	0.8009	13	长春	0.5192	24	哈尔滨	0.4321
3	昆明	0.6888	14	杭州	0.5188	25	武汉	0.4153
4	贵阳	0.6742	15	银川	0.5014	26	北京	0.4060
5	合肥	0.6310	16	南京	0.4634	27	呼和浩特	0.3738
6	南宁	0.6207	17	郑州	0.4622	28	成都	0.3712
7	重庆	0.6191	18	石家庄	0.4578	29	沈阳	0.3537
8	福州	0.6175	19	长沙	0.4552	30	乌鲁木齐	0.3223
9	太原	0.5896	20	天津	0.4534	31	兰州	0.2941
10	广州	0.5852	21	西宁	0.4520			
11	南昌	0.5798	22	济南	0.4426			

评价结果显示：①环境质量总体呈上升趋势，2008 年大幅提高，随后有所下滑（图 9-6）。2008 年各指标下降幅度较大（表 9-3），据《2008 年北京市环境状况公报》显示，这与北京召开奥运会、残奥会采取的多种污染治理措施有关。②指标值虽然呈下降趋势，但与国家标准仍有差距。2000—2012 年，除了化学需氧量和区域环境噪声分别以年均 0.34％和 0.02％增长，可吸入颗粒物、二氧化硫、二氧化氮和道路交通噪声分别以年均 3.25％、－7.45％、－2.56％和 0.21％递减，但根据《中华人民共和国环境空气质量标准 GB 3095—1996》、《中华人民共和国声环境质量标准 GB 3096—2008》和《国家环境保护模范城市考核指标及其实施细则（第六阶段）》，只有二氧化硫和二氧化氮优于国家标准，其他各指标（化学需氧量标准无法获得）均劣于国家标准（如图 9-7 所示，设定国家标准为 1）。③环境污染程度仍然较高，在 31 个城市中排在第 26 位（表 9-2）。大部分指标值标准差相对较小（表 9-4），排名靠前的指标对于综合排名提升贡献不大，北京环境质量综合得分较低，主要是大气污染所致。

表 9-3　2008 年各评价指标变化趋势

	可吸入颗粒物	二氧化硫	二氧化氮	二级以下天数	化学需氧量	区域环境噪声	道路交通噪声
年均增长率	－3.25％	－7.45％	－2.56％	－6.49％	0.34％	0.02％	－0.21％
2008 年比 2007 年增长率	－17.43％	－23.40％	－25.76％	－23.53％	－5.61％	－0.74％	－0.43％

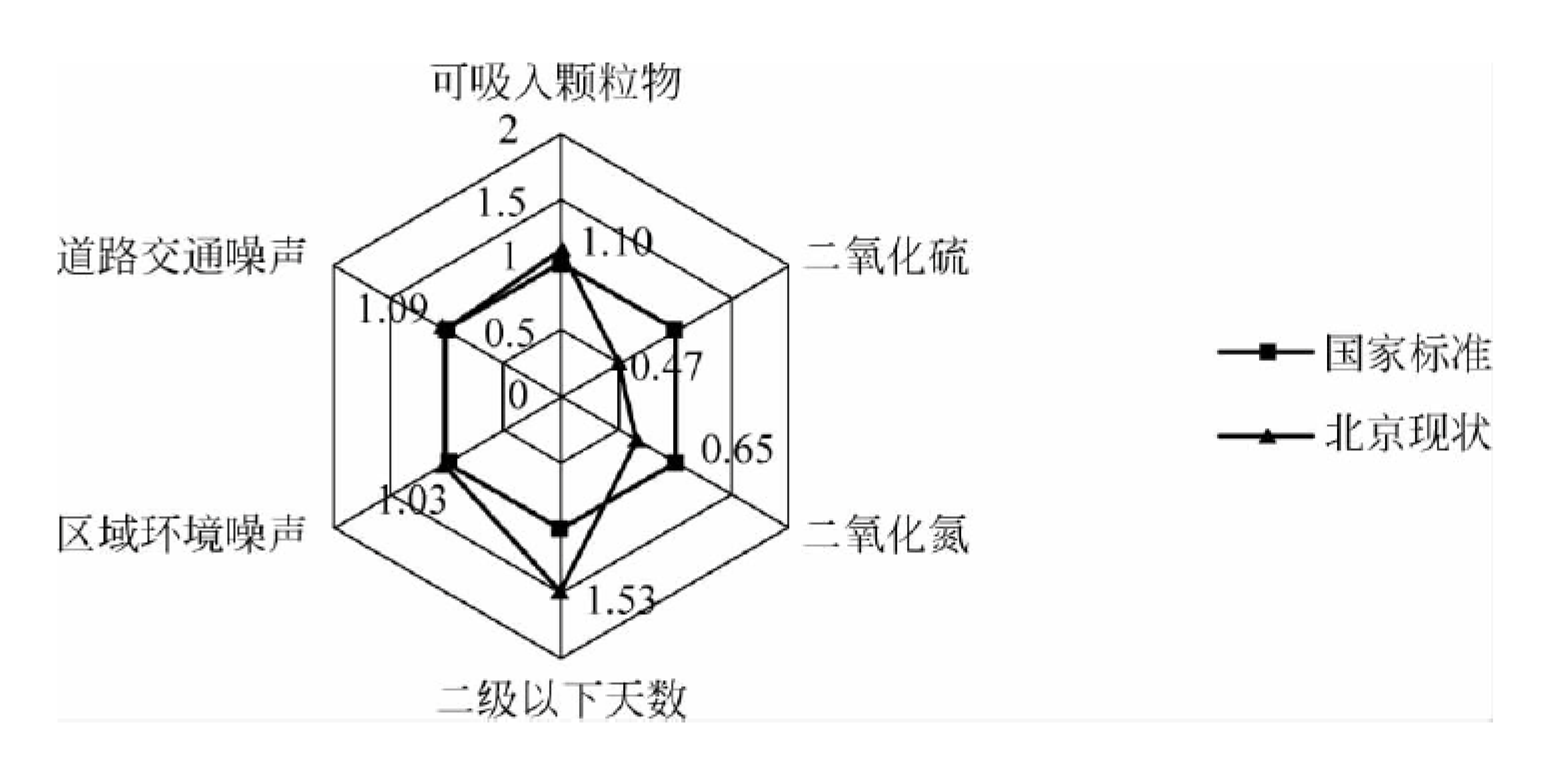
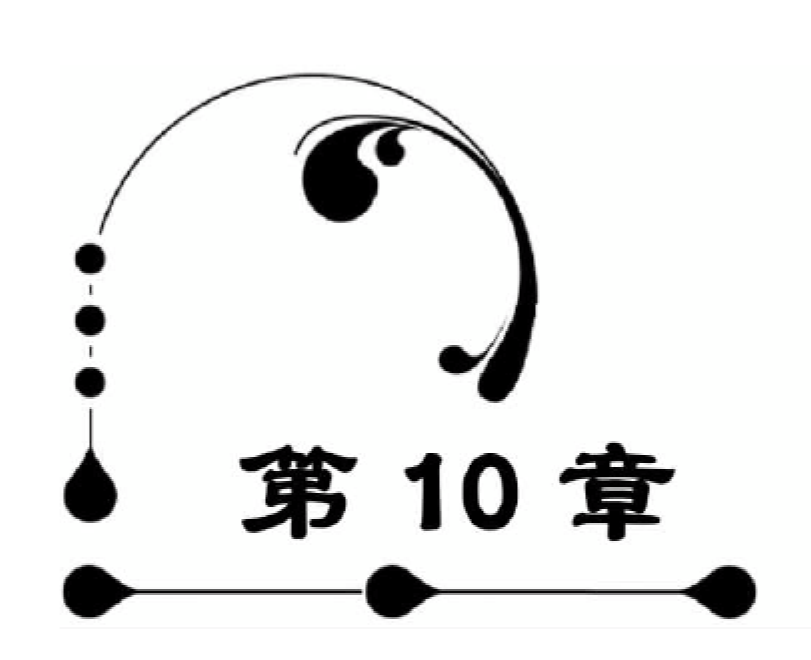


图 9-7　2012 年北京环境质量评价部分指标与国家标准对比

表 9-4　2012 年北京环境质量评价各指标 31 个城市排名和指标值标准差

	可吸入颗粒物	二氧化硫	二氧化氮	二级以下天数	化学需氧量	区域环境噪声	道路交通噪声
城市排名	27	9	28	30	4	12	21
指标值标准差	0.02	0.15	0.01	24.57	0.01	1.75	0.89



第 10 章 移动社交大数据解决方案

过去数十年的时间里，中国互联网从无到有，从弱到强，互联网产业迅速崛起。人们从一无所知到无所适从，移动社交网络在中国开始以迅猛的速度发展。中国互联网络信息中心(CNNIC)发布的报告显示，截至 2011 年 12 月月底，我国网民数量高达 5.17 亿，全年新增网民 5580 万。其中，手机网民的规模达到 3.56 亿，占网民总体比例的 69.3%。使用手机社交的用户年增长率为 35.7%，成为增长率最高的 3 个应用之一。各大社交网站在大力推进移动社交网络的同时，不断开发新的应用模式，实现了多人游戏与视频，应用服务模式不断创新，商业模式不断完善。

2012 年，随着智能手机的不断普及，移动互联网的创业高潮即将到来。由此可见，移动互联网的发展是随着时代的进步和人们对于新技术的需求而产生并不断发展的，移动互联网络深刻地改变了人们的生活。与此同时，人们对于移动社交网络的需求也促进了移动互联网络在技术与应用等各个方面的发展。

社交网络(Social Networking Services)是一种新型的网络服务，用户可以参与其中，交流、合作、分享、发布、传播信息，进而组成一种在线虚拟社区(图 10-1)。最近几年，随着 Web 2.0 的快速发展和互联网的普及，社交网络呈现爆发式增长，从早期的 UGC 内容的论坛、博客等网站到最近的 SNS、微博等新兴服务，社交网络正在成为互联网时代的新宠儿，并且极大地改变了人们获取信息和移动互联网的使用方式。



图 10-1　SNS 网络图

社交网络在提供一种在线的信息发布和传播平台的同时，也在深刻地影响着人们的现实社会。2011 年以来，中东北非的动乱充分地显示了社交网络作为一种互联网服务其影响范围已经远远超出了互联网的界限，而深入到人们的政治生活中。而不断出现在社交网络中的各种企业用户，标志着企业团体已经注意到这一新生事物，并且敏捷地试图通过社交网络中的新方式与用户交流沟通甚至达成贸易。可以预测，随着社会化浪潮的演进，社交网络一定会改进整个商业生态系统。

从我国目前的形式来看，改革开放 30 年

的发展成果使得人民的物质财富急剧增加，对于精神文化生活的需求逐渐抬头。这一趋势和互联网的普及运动相重合的结果就是社交网络这一新生事物在中国的大规模兴起。目前国外 Facebook 的用户数已经突破 9 亿，Twitter 的用户数已经达到 6 亿，国内的新浪微博和腾讯微博也都达到了 3 亿和 4 亿的规模。微博的出现，大大降低了互联网的门槛，使得发布信息、传播信息变得平等而开放，信息的传播被极大地加快，微博正在成为人的另一种生存方式。

截至目前，国内已经有几百家媒体电台和几千家各类政府和商业组织进驻，通过微博倾听民意，调查用户需求，甚至作为客户服务的主战场，都已经成为当前微博的常态。而在这种情况下，企业界的社会化营销、社会化的市场运作等新式的企业运行方式也纷纷出现。微博不仅成为人们传播信息发布信息的平台，而且成为企业获取用户、挖掘用户、维持用户、服务用户的一个平台。

在中国，截至 2010 年 8 月，已有 59 个政府部门在新浪注册政府微博，其中公安微博 40 多个。日渐兴起的“政府微博”成为政府在新时期加强和创新社会管理的有益尝试，其正不断尝试用这一新的方式搭建起与公众的互动平台，达到沟通、倾听、辅助决策的目的。2010 年在惩治腐败、城市建设、突发事件、爱心传播等许多热点问题上，中国网民都通过微博参与其中。微博的迅速发展也使其迅速成为备受各方关注的舆论新阵地。

10.1 移动社交网络发展情况

10.1.1 移动社交网络发展现状

美国“传播学之父”Wilbur Lang Schramm 认为，传播不是全部通过言词进行的，任何非语言的传播都携带着信息，而这些信息都有可能刺激所有的感官并使交流的对方同这种全身心的交流相呼应。移动社交网络就是在语言环境之外的一种非语言沟通与交流平台。它是一个开放性的社会化网络平台，能够利用移动通信设备的移动性、便捷性、及时性等各方面的优势，让用户随时随地进行沟通与交流。它能够为用户提供实时定位和信源确认等服务，增加移动社交网络平台的可信度。从目前我国移动社交网络的发展状况来看，它以现实的人际关系为基础，使用范围较为广泛，但是在使用过程中，受社交网络自身因素和人们心理因素的限制，它也存在着社交从众心理和社交疲劳等现象。

1. 用户的广泛性

麦克卢汉（Marshall McLuhan）认为“任何媒介对个人和社会的影响，都是由于新的尺度产生的，我们的任何一种延伸，都要在我们的事物中引入一种新的尺度”。与以往传统的网络传播模式比较，移动社交网络将用户的个性化需求作为新的尺度。以用户的个性化需求为蓝本，针对不同的用户，采用新的观点和尺度，建立因人而异的个性化信息表达，从而也使得其在用户群体中得到广泛应用。

移动社交网络使用范围的广泛性涵盖两个方面的含义：内容的广泛性与地域的广泛性。所谓内容的广泛性是指移动社交网络中的内容，包括内容的形式多样化，能够充分满足不同用户的多样化需求。移动社交网络充分发挥了网络媒体的资源优势与形式的优势，利用其开放性平台，在形式上涵盖了音频、视频、图片和文字等各种形式，给用户留下了深刻的印象。内容的广泛性是伴随着信息通信技术的不断完善与发展和人们对于社交网络的要求

提升而逐步发展起来的，它是移动社交网络吸引用户、留住用户所必须具备的条件之一。

所谓地域的广泛性是指移动社交网络里的人际交往和社会交往打破了时间与空间的限制，通过各种移动设备，能够在较广的范围里与他人保持沟通与联系。移动社交网络最大的优势在于其移动性可以使人们在内容、地域等方面达到较为广泛的接触面，能够消解物理中介所产生的差距，达到随时随地的“人-机”互动和“人-人”互动，消解空间距离。移动社交网络的广泛使用表明它已经扩散到大众中间，具有浓厚的草根气息。“草根性具有强大的凝聚力，更具有强大的生命力和独立性”。移动社交网络为人们创立了一个平等、自由的话语空间，利用其便捷的转发与对话功能，可以使更多的人能够通过移动终端与他人保持联系，提供最新的信息，创作丰富的内容，同时以更快的速度传向世界各地。

2. 用户易产生从众心理

在这个信息爆炸的时代，人们接收与认知信息的渠道不断增多，各种正面、负面的信息铺天盖地，每一种言论都代表一方的利益和态度，对于一些信息，人们无从判断真假与对错，也无法全面而准确地认知各种社会事件，所以在一定程度上也只能参考一些他人信息，从众心理和行为便应运而生。它是信息社会的产物，但也是人们对未知事物进行判断的一种客观需求。在移动社交网络的发展中，用户在选择和发表言论等方面都存在一定的从众心理。

(1) 选择的从众性。在人类社会向现代化迈进的过程中，现代社会的紧张与压力拉远了人际之间的距离，人们渴望自由和相对独立的空间，希望自己和哪怕是最亲近的人之间都保持一定的距离，人与人之间的关系疏远，人际交往淡化，人们最终被淹没在大众传播的浪潮里。但是，很多人并不甘于被湮没，正如社会心理学家戈夫曼所言，社会就像一个舞台，人人都是演员，每个人都在有意或无意地通过自我表现给别人留下印象。所以，有时为了与他人保持一致，与社会保持一致，人们在选择时，便会较多地考虑你是怎么想，他会怎么看，一般都会采取与社会大多数人相同或相似的态度，而不会选择与社会背道而驰。因此，选择的从众心理自然而然便会产生。

(2) 言论的相似性。用户的言论是受其认知和情感因素影响的，从理论上来看，不同用户对于同一事物的言论是存在差异的。卡罗尔·E. 伊萨德(Carroll E. Izard)认为，想法或态度是一种时间有限的特定的情感过程，从几秒到几小时，从温和到激烈都有可能。除此之外，人们还会表现出一些情感的特征，也就是在与他们的接触中倾向于表现出的某种特定的情感。在移动社交网络中。受时效性和移动设备的限制，用户在与人交往的过程中，往往较少采用理性的头脑去分析和判断，对于平台中的信息只是简单的过滤，并没有进行深入的分析和思考，这使得发表的言论比较浅显，而为了不被别人发现自己的言论比较浅显，很多人会随别人的意见和态度而发表意见，并不完全体现真实的自我思维与态度。这种较为关注自身形象的信息传播方式造成用户不愿主动表达出自我真实想法，而是与社会大众的态度保持一致。因此，言论的从众心理也自然就出现了。从众心理是一种较为普遍的心理和行为，这主要是因为人们在自我意识方面存在一定的弱化现象，在思维方面缺乏独立思考的能力。个体意识不强。在面对意识判断与抉择的过程中存在犹豫不决的现象和徘徊心理，无法根据内心最真实的想法做出理性客观的判断，这不仅会影响社交网络的有效性，同时也在一定程度上会影响用户参与社交网络交流的积极性与主动性。所以在社交网络发展的过程中，应该尽量避免出现用户的从众现象，尽可能鼓励用户发表内心真实的想法，自主表达言论。

(3) 用户易产生社交疲劳。人们除了需要应对现实社会中的难题之外,同时还要应对网络中层出不穷的各种问题,在人际交往方面也是如此。人们在通过移动互联网进行人际关系的开拓与维系外,同时也要面对现实社会中人际关系的建立与维系,这都需要花费大量的时间与精力,用户易产生社交疲劳。

埃瑟·戴森说:“数字化世界是一片崭新的疆土,可以释放出难以形容的生产能量,但它也可能成为恐怖主义和江湖巨骗的工具,或是弥天大谎和恶意中伤的大本营。”作为一种人际沟通工具,移动社交网络的影响力是众所周知的,它对人们的生活方式及思维方式都产生了较大的影响,在为人们提供各种便利的同时,也会被各种垃圾信息、虚假信息所埋没,严重影响了人们获取信息的信度与效度,使人们容易产生视觉疲劳,人们对于社交网络的信任程度与依赖程度降低。另外,移动社交网络中过多的利益导向性使得社交网络中除了自己需要的信息外,还有较多冗杂的内容影响个人需求的满足,用户可能需要较多的时间才能达到维系人际交往的目的,无法使预期效果最有效实现,使用户产生疲惫的感觉,疏于移动式的人际交往。

从移动社交网络的发展状况来看,虽说移动SNS用户人数仍呈增长趋势,但由于内容的易用性和丰富性不够、移动网络的性能不稳定、应用收费模式设置不合理等原因,移动社交网络面临较大的发展瓶颈。所以移动社交网络的发展必须结合中国的特色和用户的实际需求,打破原有传统的限制与约束,力求创新,只有这样才能使分散的社交网络用户进一步集中,得到较快的发展。

10.1.2 移动社交网络发展方向

我国的移动社交网络要想走得更远、走得更快,就必须进行资源整合。移动社交网络的整个设计、应用、商业模式等,结合中国的文化要素,进行创新性设计,突出文化特色。目前我国的移动社交网络在发展过程中大多是模仿国外社交网络的架构、应用等进行操作,虽有一些创新性的应用方式和商业模式,但从本质上和长远发展的角度来看,它缺少文化力。国外的移动社交网络之所以能够得到较快的发展,很重要的一个因素是它是应时代和公众的需求而产生的,它体现了人们对于社会交往的一种需求心态,有强大的文化力。此外,应用设计并不是越新或越全面越好,最重要的是要与用户的心理特征和社会交往的需求相结合,要为用户提供其最需要的社会交往信息,并以最为快捷的方式展现,只有这样才能确保社会公众对移动社交网络的认可与关注。

移动社交网络的发展要掌握公众的媒介动机和社交的需求。“使用与满足”理论认为,用户在接触媒介的过程中有一定的需求动机或目的,希望从中得到满足;由于社会环境和媒介是不断变化的,人们的动机也是不断变化的,需要不同的内容在不同的方面使用户得到满足。因为用户环境、媒介内容等整体的社会背景影响用户媒介行为,用户媒介行为的形成与其社会状况与需求、大众媒介结构产品有较大的联系,只有用户有媒介需求和动机,社会有对应的媒介结构与产品时才能产生用户的媒介行为。用户媒介行为又可以对他们给予反馈,进而促进媒介平台的发展。任何一种网络平台或媒介平台存在的依据是市场需求,存在的目的是满足用户需求,而非简单体现平台的技术与应用等。现代认知心理学认为,人们的行为并不是一个对外部刺激做出的纯粹被动的反应。主体的选择、加工在受众与大众传媒之间起着十分重要的作用,这种作用的发挥与受众个体的认知结构密切相关。因此人们获

得信息的过程在于新的信息与主体已有认知之间的相互联系与作用，两者的互动决定着现实生活和信息传播活动中人们学习过程的本质，这一本质蕴涵在主体认知结构的不断扩展、分化和重组的过程中，而认知结构本身正是通过这一过程得到更新，从而为人们进一步的认知实践提供了新的基础。伴随着移动互联网的快速发展，移动社交网络的发展并不能再仅仅局限于原有的商业模式和应用平台，它必须与社会环境和个人需求进行全面的结合，在准确认知社会客观媒体环境与发展环境的基础上，及时掌握公众的媒介动机和社会交往的需求，只有这样才能使得移动社交网络不断创新，在竞争中处于优势地位，获得全面发展。在互联网和物联网快速发展的今天，移动社交网络逐步向人工智能的方式转变，人们通过移动社交网络可以根据自身此时、此地、此物的切身愿望反馈和获取信息。如图 10-2 所示为未来移动社交网络示意图。

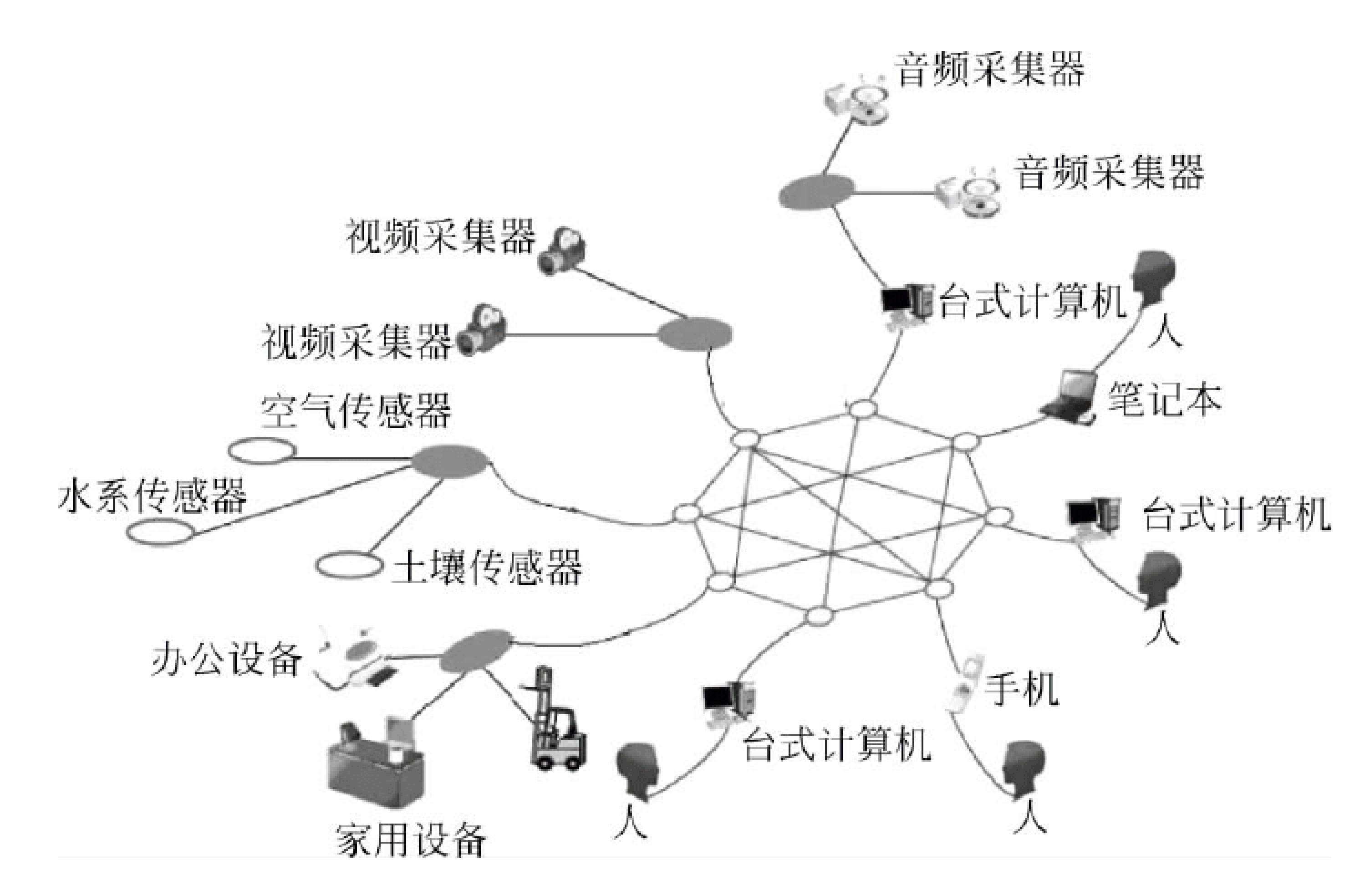


图 10-2　移动社交网络示意图

10.2　社交网络基础理论和商业模式

10.2.1　社交网络相关理论

1. “六度分隔”理论

1967 年，美国哈佛大学的心理学教授 Stanley Milgram(1933—1984)想要描绘一个连接人与社区的人际联系网，做过一次连锁信实验：他将一封信件随机寄给了位于美国中西部内布拉斯加州的 160 个人，信中印有千里之外波士顿的一名普通股票经纪人的名字，米尔格拉姆在信中要求收信人将这封信通过自己的朋友寄给收信人，结果大多数人只经过了五六个步骤，这封信就最终到达了这个股票经纪人的手中。结果发现了“六度分隔”现象。六度分隔现象(又称为“小世界现象”)，可通俗地阐述为：“你和任何一个陌生人之间所间隔的人不会超过 6 个，也就是说，最多通过 6 个人你就能够认识任何一个陌生人。”

“六度分隔”说明了社会中普遍存在的“弱纽带”，但是却发挥着非常强大的作用。有很多人在找工作时会体会到这种弱纽带的效果。通过弱纽带人与人之间的距离变得非常“相近”。这个理论在社交网络中也被广泛应用，最典型的是扎克伯格(Mark Zuckerberg)创建的 Facebook 网络产品。通过同学圈、朋友圈、社会圈、同事圈等可以在 6 个人内找到要找的

人，绝对没有联系的 A 与 B 是不存在的。网络更加缩短了空间、拉近了距离。

2. 弱关系、强关系

马克·格拉诺维特在 1973 年发表的论文中指出：在传统社会，每个人接触最频繁的是自己的亲人、同学、朋友、同事等，这是一种十分稳定的然而范围有限的社会关系，这是一种“强关系”；同时，还存在另外一类相对于前一种社会关系较浅，然而却是更为广泛的社会关系，格兰诺维特把后者称为“弱关系”。

研究发现：其实与一个人的工作和事业关系最密切的社会关系并不是“强关系”，而常常是“弱关系”。“弱关系”虽然不如“强关系”那样坚固（金字塔），却有着极快的、可能具有低成本和高效能的传播效率。

事实上，在信息的扩散传播方面，“弱关系”起着同样的作用。一个人的亲朋好友圈子里的人可能相互认识，因此，在这样的圈子中，他人提供的交流信息重复度高。比如，我从这个朋友或亲戚听到的，可能早已经在另一个朋友那里听说了，而他们之间也都相互交谈过此话题。日常生活中不乏这样的事例。

弱关系在我们与外界交流时发挥了关键的作用，为了得到新的信息，必须充分发挥弱关系的作用。这些弱关系，或是熟人，都是我们与外界沟通的桥梁，不同地方的人通过弱关系可以得到不同的信息。最亲近的朋友可能生活圈子和你差不多，你们的生活几乎完全重合。而那些久不见面的人，他们可能掌握了很多你并不了解的情况。只有这些“微弱关系”的存在，信息才能在不同的圈子中流传。弱关系的威力正在于此。

强连接关系通常表明行动者彼此之间具有高度的互动，在某些存在的互动关系形态上较亲密，因此，透过强关系所产生的信息通常是重复的，容易自成一个封闭的系统。网络内的成员由于具有相似的态度，高度的互动频率通常会强化原本认知的观点而降低了与其他观点的融合，故认为在组织中强关系网络并不是一个可以提供创新机会的渠道。

事实上，强弱关系并不仅由人与人之间的关系类型决定，还会由六度理论的度数决定。可以理解的是：1 度关系肯定要比 2 度关系强。此外，如果在 SNS 中，强弱关系还可能会根据建立关系的依据来决定，同爱好/同兴趣、同群组/同圈子、同应用，这类关系相对较弱，但同一类关系的交集越多，关系则可能会越强。

3. 贝肯数

贝肯数是基于“六度分割”理论演进而来的。贝肯是好莱坞的一名普通演员，不同于马龙·白兰度这样的大腕，贝肯在好莱坞电影中从来都是以配角的身份出现，他与当时好莱坞的影视明星发生联系所需要的中间人数量即为“贝肯数”。弗吉尼亚大学一个实验室曾为约 25 万上过银幕的男女演员计算了他们的“平均贝肯数”，研究发现，无论是历史上贝肯数最低的演员罗德·斯泰格尔，还是一个名不见经传的小演员，他们的贝肯数都在 2.6～3 之间，并且相差十分微小。

这一发现说明，其实你要想进入网络的中心，并不一定要成为大人物，即使成为一个“永不退场”的配角也可以非常接近网络的中心，你和中心人物的距离其实可以近到忽略不计，因为那不是一个物理距离，而只是一个连接度的问题。“贝肯数”的发现还说明要想阻断一个网络和另一个网络的连接（比如让马龙·白兰度永远和某个导演无法接触到），隔离“贝肯”这样的高连接性人物就可以了。同样，一个网络社区的崩溃，其实不会因为多少普通用

户流失而发生，但几个节点用户的流失，就会造成崩溃。有趣的是，“贝肯”在哈佛大学的学生中被当作一种“比拼记人名”的游戏，即背出和“贝肯”合作过的明星，当然这个游戏也可以把“贝肯”换成其他领域里的某个高连接者。

4. 顿巴数

“顿巴数”是英国牛津大学人类学教授罗宾·顿巴(Robin Dunbar)在 1992 年的一项研究成果。

根据顿巴教授的研究，人类的社会结构表现为：5 人左右的亲密接触圈；12～15 人的同情圈，即，如果这一圈里有人痛苦，我们也会伤心；50 人左右的群落，即经常一起生活、一起行动的人(已经有限定在这一人数内的社交网络工具出现)；150 人左右的氏族，即遵从共同仪式的人；500 人左右的部落，即拥有同种语言的人(其实在现代社会，这里的语言有时只是指一些经常交流的人之间约定俗成的词语和概念，外人第一次听到不能理解)；5000 人左右的群落，即有共同文化的人。按照顿巴数的同心圆模型，当社会结构的人数超过 150 人时，相互间的互动和影响就会减少很多，只能靠共同的语言来维系，而当人数上升到 5000 人左右时，维系社会结构则只能依靠共同的文化。

10.2.2 社交化商业模式

互联网是虚拟世界和现实世界的桥梁，在互联网上将现实生活中人与人之间的关系建立起来。互联网的发展为社交网络的发展奠定了基础，社交网络的发展同时也让互联网的关系网越来越复杂，在这种情况下，Facebook 创始人马克·扎克伯格提出了社交图谱的概念，因此，也让他的网站一举成名。不论是国外的 Facebook、Twitter；还是中国的微博、开心网、微信等基于互联网的社交网络已经深入人心，IT 产业步入以社交网络为主的关键时刻。

社交网络的结构主要由以下 4 个方面组成：用户、内容、社会网络和工具。这 4 个方面相辅相成、彼此依赖，但不可否认，以数据为载体的内容是这个结构的核心。这是因为用户因内容分享而连接，工具因内容传播而存在，网络因内容众多而产生。社交网络每天吸引着数亿的用户在各个社交网络平台上发布自己的状态：心情、位置、爱好等。通过对这些规模化的海量大数据分析，可以从不同用户分类、用户行为以及人际关系方面获得用户规律和预测分析。通过这些用户行为分析，可以与用户之间进行良好互动，也可以为用户提供很多需要的信息和服务。企业必须重视并思考这种全新的互动方式带来的积极意义和无限机遇。

哈佛商学院副教授安德鲁·麦卡菲于 2006 年首先提出“企业 2.0”概念。他认为，企业 2.0 是企业自发性社会化软件平台，或者企业与其客户、合作伙伴及供应商之间的自发性社会化软件平台，社会化软件正使“人机交互”变成“人人交互”，企业管理也在从“以流程为中心”向“以人为中心”转变。

管理大师加里·哈默率先提出，利用互联网技术催化组织，产生管理变革；他坚信，以互联网革命的契机，将会衍生出 21 世纪新型管理模式，那就是“企业 2.0”，它是以人为本，真正尊重、激发与赞赏人的创造性、激情和勇气，以企业员工为核心，自动自发的内在需求，共同分享知识，协同合作。

“企业 2.0”提出社会化商业新生态模式，将通过企业社交网络的信息中心根据用户的不同特征构建相应的社交圈。一方面，构建企业内部社会化沟通与协作的内部社交圈，帮助

企业实现新型办公协同与知识管理；另一方面，构建企业、合作伙伴和客户关系协同社交圈以及企业通过电子商务进行营销的外部社交圈。建立以人为中心的社交管理模式，推动组织与文化变革，激发企业创新力，催生企业商业新生态。

10.3 移动社交网络数据处理架构

早期的社交网络出现在 2000 年之前，主要以 BBS(Bulletin Board System，电子公告牌系统)、新闻组等方式出现。2002 年，随着 MySpace、Facebook 的兴起，SNS 逐渐流行。国际上流行的主要社交网站有 Facebook、Twitter、LinkedIn、Google＋等；国内也有人人网、开心网、微博、微信等。他们的应用处理架构基本相似，以下将详细举例说明。

10.3.1 移动社交网络服务架构模型

SNS 产品核心结构分为 3 层，自上而下分别是：用户层(Customer)，即“用户属性和行为描述”；社区层(Community)，即“用户群内部关系链”；内容层(Content)，即“内容和应用”。“内容和应用”主要包括：官方内容；用户插件；UGC(用户生成信息)、互动游戏、群体行为、个人应用、Feed 等。“用户群内部关系链”主要包括：用户关系、群关系、关系维度等。“用户属性和行为描述”主要包括：基本用户属性、扩展属性和社区属性。具体到一个社区产品模型，从下到上也分为 3 层，如图 10-3 所示。

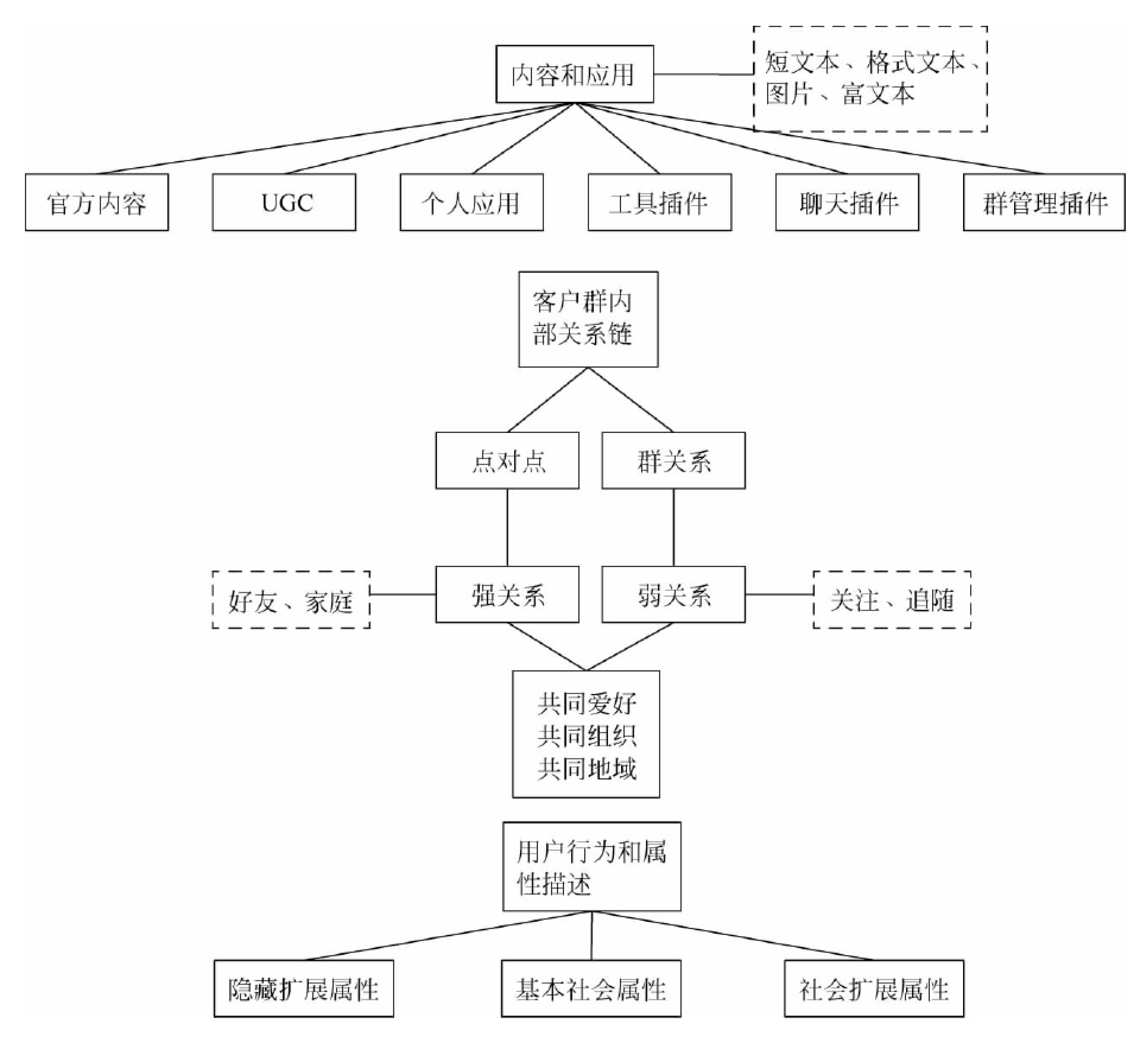


图 10-3 SNS 三层结构模型

1. 底层——用户属性描述及行为画像

用户属性比如社会属性，姓名、性别、年龄、职业、爱好等属性，还包括服务使用倾向等指导属性。

用户属性共分为3类：一类是用户的直接属性，一类是用户在社区生态中生存获得的属性，还有一类是用户隐藏的拓展属性。

用户属性分类如下。

直接属性：姓名、年龄、性别、职业、学校、毕业年份等。

生态属性：成长等级、称号、虚拟职务与相应的权利、角色等。

拓展属性：对用户的生存数据进行挖掘分析，推测出的适用于用户的个性化推荐属性。

2. 中间层——用户群内部关系链

用户群内部关系链包括人与人的关系、人与群体的关系、群体与群体的关系。具体表现为：好友关系(强关系链)、关注关系(弱关系链)、同好关系(同专业、同爱好、粉丝)、同事关系、同地域关系等。

由于用户群中的关系链连接了家庭、朋友、同事、同学、亲戚、社群等来自不同资源的群体，彼此的信息交流和互动与个人的工作生活等发生着密切的关系。由于智能手机的快速发展、用户群关系的建立更加便捷、高效，如微信、微博、社群APP等。

3. 顶层——内容和应用

(1) 内容分类包括官方发布信息，比如咨询、图片、音乐、视频等官方制作的内容；还有用户个人信息，如个人博客、即时短文、音频、视频、照片等由用户自己制作的内容。

(2) 应用。随着移动互联网的发展，APP可以表现一个互动游戏或应用软件以插件的方式与SNS轻度耦合，在移动的平台上独立运行。SNS中的APP，要调用到底层用户属性信息和中间层关系链信息以及电子支付信息等。

10.3.2 Facebook应用案例

Facebook是一个起源于美国的虚拟化社交网络服务网站，于美国时间2004年2月4日下午3点上线。截至2012年9月，Facebook拥有超过10亿活跃用户，累积了11 300亿个Likes，照片则超越2190亿张，其中有170亿张有地点信息用户可以创建个人专页，添加其他用户作为朋友并交换信息，包括自动更新及实时通知对方等。

1. 产品架构

对于http://fettermansbooks.com和http://facebook.com的共同用户来说，此时Internet应用的图景如图10-4所示。

在一般的n层架构中，应用将输入(对于Web来说，就是GET、POST和Cookie信息的集合)映射为对原始数据的请求，这些原始数据可能存在于数据库中。它们被转换为内存中的数据，并通过一些业务逻辑进行智能化处理。输出模块将针对显示对这些数据对象进行转换，变成HTML、JavaScript、CSS等。这里，在图的顶部，是运行在基础设施之上的应用程序n层栈。在应用出现在Facebook平台之前，Facebook完全运行在同样的架构上。重要的是，在两个架构中，业务逻辑(包括Facebook的隐私)实际上都是根据一些规则来执行的，这些规则建立在系统的某些数据组件之上。

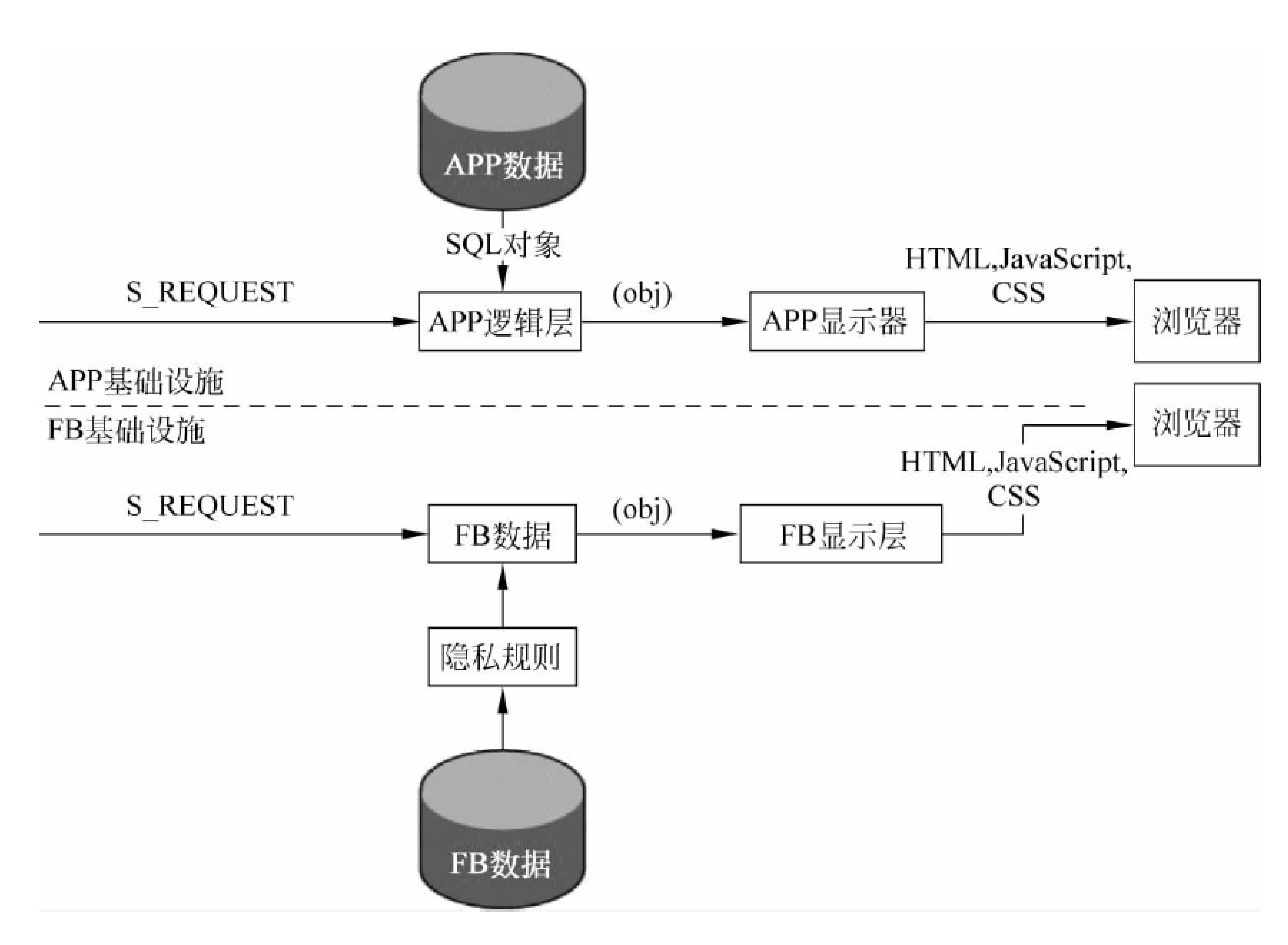


图 10-4　分离的 Facebook 和 n 层应用栈

更大量的相关数据意味着业务逻辑可以提供更多个人定制的内容，所以在 http://fettermansbooks.com（或其他应用）上浏览书籍、写书评、阅读或购买的体验，会被来自 Facebook 的用户社会关系数据加强和放大。具体来说，显示朋友的书评、期望清单和购买情况将有助于用户的购买决定，发现新的书籍，或强化与其他用户之间的联系。如果 Facebook 的内部映射 user_get_friends 可以由 http://fettermansbooks.com 这样的其他外部应用访问，就会为这些原本分离的应用提供强大的社会关系上下文，让应用程序不需要创建它自己的社会关系网络。所有这种类型的应用都可以与这种数据进行很好的集成，因为开发者可以将这些核心 Facebook 映射应用于无数其他 Web 应用，用户在这些应用里提供或消费内容。

2. 数据存储

在 Facebook 中，数据层采用了多种存储系统，包括：

（1）MySQL；

（2）Memcached；

（3）HBase（NoSQL）；

（4）Hystack（for BLOBs）。

MySQL 和 HBase 前面已经详细介绍过，分别是 SQL 和 NoSQL 数据库。Memcached 是一个流程的缓存，被用作 MySQL 缓存。Hystack 是 Facebook 开发的一个大对象存储，用来存储照片、音频、视频、邮件附件等，却不会修改文件。这些文件在 Facebook 已经有了 100PB，每天有 2.5 亿张照片上传到 Hystack 中去。随着数据量的增大，Facebook 在存储技术中逐步改善性能，由于原来的目录结构访问需要耗时长，改进了 NFS 的 Handler 缓存，减少了输入输出次数，耗时是原来的 1/3。同时还开发了 Haystack，可以将若干图片拼成一个大文件，索引放到内存中，磁盘可以一下定位到图片。读取图片只需要一次输入输出操作。

在数据分析上，Facebook是Hadoop的重要使用者和共享者。图10-5展示了Facebook的大数据分析系统。Facebook通过HBase和Hadoop来处理实时数据。由于MapReduce随机读取性能差，因此大量使用了HBase。HBase是一个高性能、高可靠性、面向列、可伸缩的分布式存储系统。利用HBase可以在廉价的PC Server上搭建起大规模结构化存储集群。Facebook还开发了一个名为Puma的流聚合处理引擎。如图10-6所示为Puma的流聚合引擎架构。

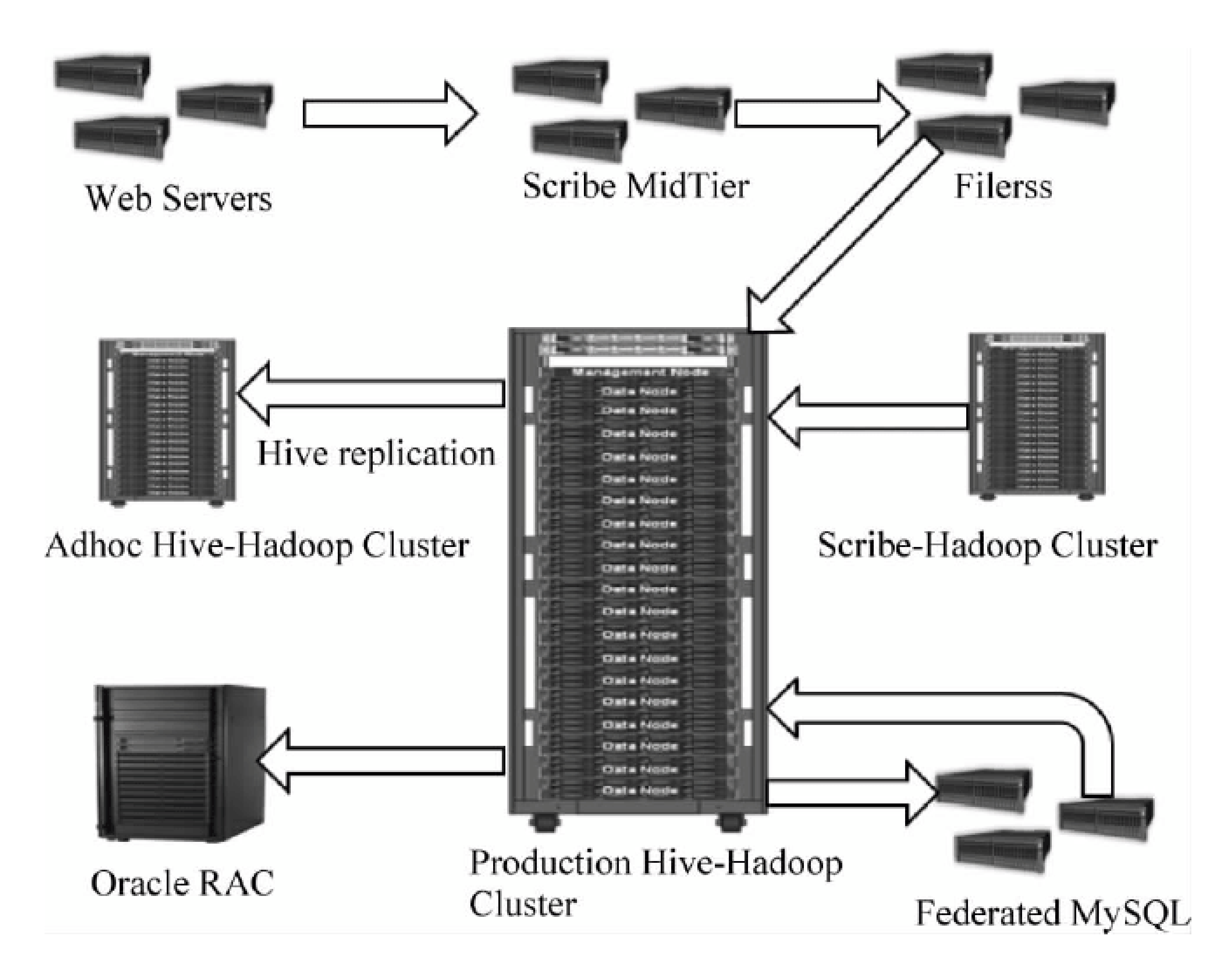


图10-5 Facebook大数据分析系统

PTail是将数据从文件系统转换成流的办法。使用检查点来通知。数据流被转发到Puma中。Puma是根据Aggregation Key分区的，每个分区在内存中保存一部分数据，并将数据持久保存到HBase中。和一般的流处理引擎不同的是，Puma还可以从HBase读取数据，来完成Join操作。Facebook公司预计将实时数据处理能力从10秒多缩减到5秒左右，大幅提升处理性能。

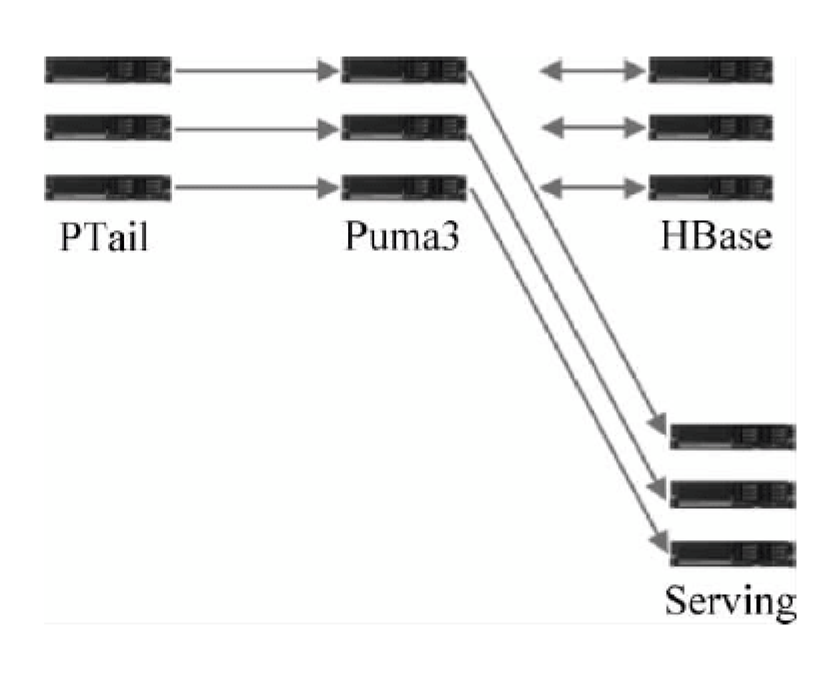


图10-6 Puma流聚合引擎架构

3. Facebook信息推荐

用户通过SNS平台可以自主产生内容，包括对内容进行浏览、收藏、转发、分享、评论、编辑等各类操作产生的数据，以及用户与用户之间通过关注、加好友等方式留下的大量的、即时的、多样化的数据。通过利用数据挖掘技术对这些数据进行挖掘分析，可以立体地勾勒出用户的影像。Facebook在做这些研究时，通常会为所有的内容进行加权，并最终计算出用户对内容的喜好程度以及用户与用户之间的关系权重。这样的好处就是可以更加精准地为用户推送个性化内容，以及可以很好地测量出用户与用户之间的紧密程度。现在对于每个用户而言，每天通过SNS接收的信息会非常多，甚至感觉到有些顾及不了。通过大数据分析，如果可以锁定用户关心的知识范围和喜好，就可以通过SNS直接推送，提高用户的阅读信息量，同时也节约了用户阅读时间，久而久之，对每个用户的受益会更大。如图10-7所

示为用户浏览信息大数据分析图。

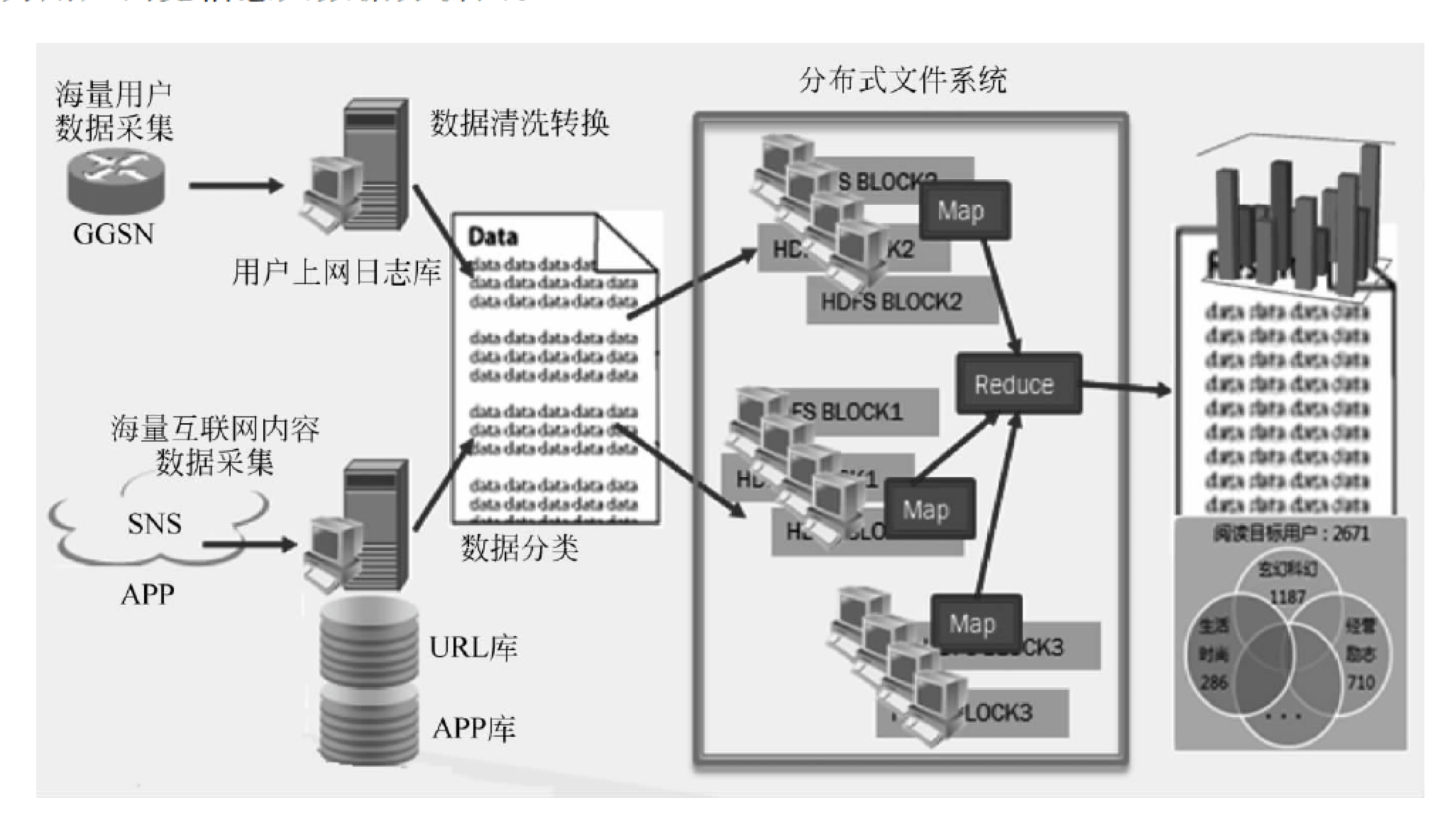


图 10-7　用户浏览信息大数据分析图

4. Facebook 消息订阅

New Feeds 信息流，指网站发布的所有最新内容列表。用户可通过对目标网站上的 News Feeds 进行订阅，接收新发布的内容，如图 10-8 所示。New Feeds 对 Facebook 的发展而言是功不可没的，对整个 SNS 行业发展也起到了重要作用。New Feeds 并不是 Facebook 发明的，却很好地应用在社交网络中，并对信息沟通与信息传递提供了巨大的支撑。随着互联网行业的快速发展，用户对个性化内容的诉求越来越强烈，RSS 应运而生。RSS 约定了一种信息共享方式和数据格式规范；用户可以事先设定好过滤条件，信息有更新时，主动从 New Feeds 信息源 PUSH 到用户面前。随着时间的推移和技术的成熟，必将得到更加广泛的应用。

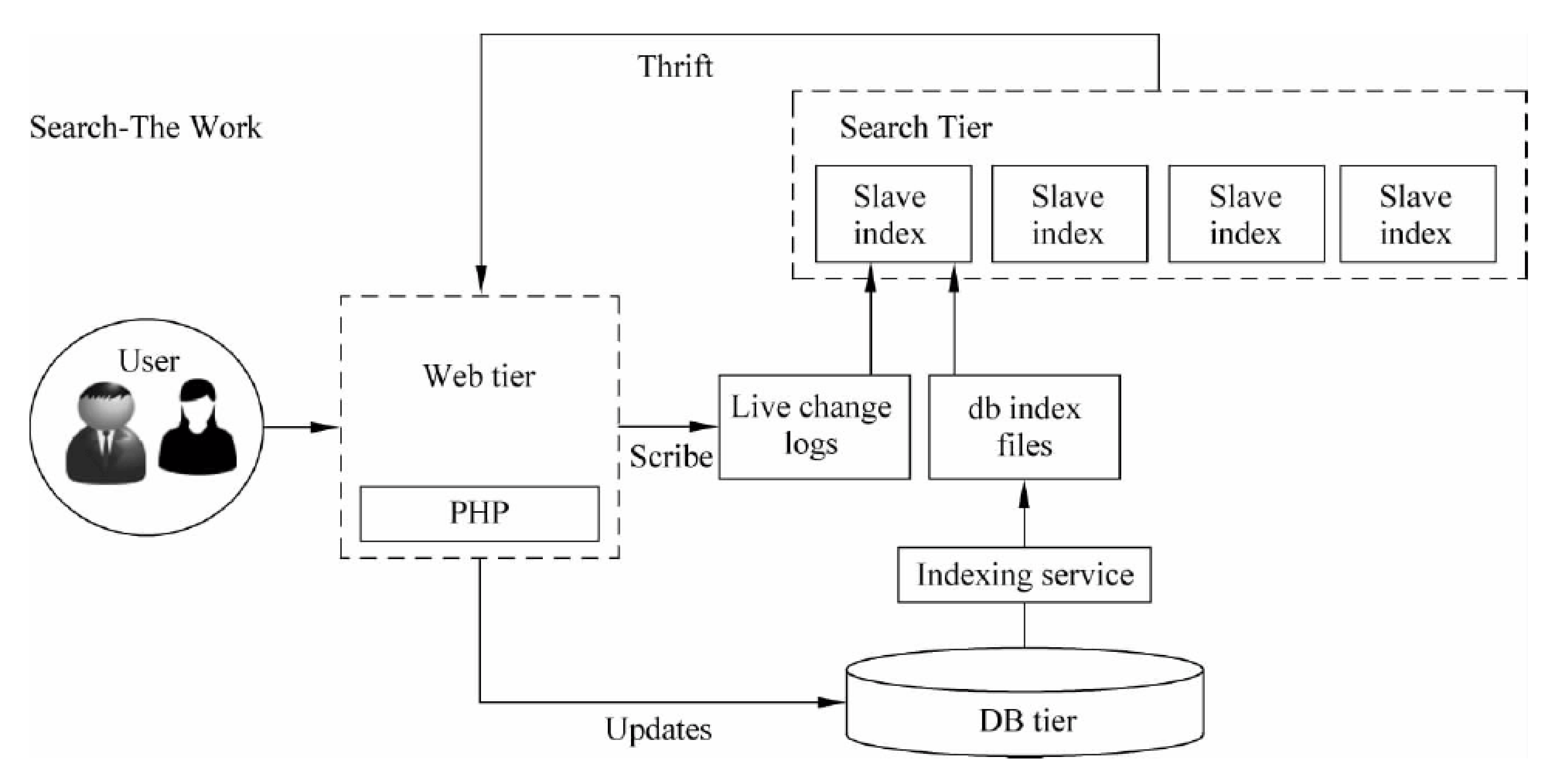


图 10-8　Facebook Search 的架构示意图

随着用户之间关系的密切发展，信息动态在关系链之间的传播非常重要。用 New Feeds 来处理此类问题是非常适合的方案。在 SNS 平台上，用户可以默认将自己的动态通过 New Feeds 传播到关系链中，也可以等待朋友的最新动态。

10.4 移动社交网络大数据分析

事实上，随着 Facebook 等社交网站的出现与普及，企业界很快就出现了以市场应用为导向的影响力分析机构。但是由于普遍没有学术界理论的指导，大部分影响力算法都是经验总结式，并没有提出明确的评价方法。下面择要进行阐述。

(1) Klout 是美国一家专注于评价用户在社交网络中的影响力的数据挖掘公司。目前已经可以追踪用户在 Twitter、Facebook、LinkedIn 等多家社交网站上的数据，以此计算用户的影响力大小。

随着社交网站的崛起，Klout 也在人们的现实生活中发挥了巨大的作用。2012 年 2 月，企业软件公司 Salesforce 开始引入一种新服务：让客户公司可以根据投诉客户的 Klout 打分来灵活处理投诉。对于那些影响力大的客户，其投诉将会处理得更快，获得的服务也会更好。一家著名的奢侈品购物网站 Gilt Groupe 也开发了一个新的产品：根据客户 Klout 的打分提供不同的折扣。Klout 计算用户在某个社交网络中的影响力的算法简述如下：真实覆盖度(True Reach)：你可以影响多少人？系统过滤掉了那些垃圾用户和僵尸粉，着眼于那些可以被你发布的内容引起动作的用户。

① 扩散概率(Amplification)：那些人受你的影响有多大？这一指标主要受到转发你消息的人数和他们转发你消息的频率的影响。

② 网络影响力(Network Impact)：你所在的网络影响力有多少？这一指标主要考虑你所在的社交网络中是否包括某些影响力高的人。

Klout 倾向于评价一种绝对影响力(全局影响力)，后来引入了主题这个概念，但是是用类似人本计算的方式来票选出某个主题上影响力最大的人物。

(2) Kred 是一家新兴的社交网络影响力分析企业。Kred 基于以下两个标准衡量用户的社交影响力。

① 影响力：能够激起他人行为的能力，比如别人的转发、回复等。

② 扩展力：人们试图激起你行为的尝试，比如与别人的交互、他人赠送的礼物等。

相比 Klout，Kred 做得更加精准。Kred 基于社区来识别用户的影响力。其建立了一种按照社区对不同话题的集合影响深度。他认为，“由真实用户所组成的紧密的小圈子，才是影响力的摇滚明星”。Kred 会根据你的简介数据定义社区，然后为这个社区计算一个集合影响广度和深度的分数。它还会计算你在你某个社交圈中的排名。

10.4.1 社交网络平台行为影响分析模型

社交网络用户影响力分析概率模型框架——一般阈值模型。定义用户被信息影响或者称该用户被激活，就会成为传播节点，将该行为在社交网络中继续传播下去。某一时刻未被影响的用户节点 u 周围已经有若干父传播节点，他们形成对用户 u 的影响用户集合 S，集合中任意用户节点 $V \in S$ 都是在用户 u 关注用户 v 之后被激活的。影响用户集合 S 中的任意

用户 v 均会以一定的概率激活用户 u,从而集合 S 中的所有用户会形成影响联合概率 $P_u(S)$ 而对用户 u 产生影响,用户 u 被影响后就会发出同样的行为。

本节主要讨论计算影响用户集合 S 对用户 u 的影响联合概率,对单个用户 u 的发出行为做出预测,一般阈值模型中的阈值 θ_u 指用户 u 的受影响阈值,当 $P_u(S)\geqslant\theta_u$ 时,可以预测用户 u 将会被父节点影响,从而成为传播节点,获得传染性,可以将行为继续传播到子节点中未被影响的用户节点。

根据微博中用户间的实际影响关系,显然可知,影响联合概率函数 $P_u(S)$ 是单调的,如果 $S\subseteq T$,一定有 $P_u(S)\leqslant P_u(T)$。而且用户 u 的所有父节点之间具有比较弱的联系,父节点对用户 u 的影响概率可以看作是独立的。

因此,影响联合概率可以被定义为

$$P_u(S) = 1 - \coprod_{v\in S}(1 - P_{v,u}) \tag{10-1}$$

式(10-1)中,$P_{u,v}$ 指用户 v 对用户 u 的行为影响概率,也就是行为从用户 v 传播到用户 u 的概率。行为传播有延迟时间,用户 v 发出行为 a 后,用户 u 在 $t_{v,u}$ 的延迟时间后被影响而发出同样的 a 行为。$t_{v,u}$ 是通过历史微博记录统计出的行为从用户 v 传播到用户 u 的平均延迟时间。用户 v 和用户 u 之间行为传播平均延迟时间定义如式(2)所示

$$t_{v,u} = \frac{\sum_{a\in A}(t_u(a) - t_v(a))}{A_{v2u}} \tag{10-2}$$

式(10-2)中,$t_u(a)$ 表示用户 u 发出行为 a 的时间,A 表示微博中的历史行为集合。

上述公式中,假设任意用户的所有父节点对该用户的影响都是独立的,父节点对该用户的影响没有依赖关系,因此,如果能计算任一父节点对该用户的影响概率,就可以通过式(10-1)计算该用户受到所有父节点的行为影响联合概率,即该用户发出同样行为的概率。但在实际社交网络中,还要考虑动态性,用户之间的影响概率应该是一个连续时间函数。

10.4.2 社交网络单平台内影响力分析

美国的一家创业公司 Klout,2009 年开始研究推特(Twitter)用户影响力指数,2010 年开始把影响力测量产品推向脸书(Facebook)。Klout 指数(Klout Score)用于表征用户在推特和脸书上的综合影响力,这项指标介于[1,100]区间,反映了用户在推特和脸书上行为的 35 个变量。图 10-9 显示了网友 Cristen Perks 的 Klout 指数和其品牌。具体算法涉及 3 个因素:反映粉丝质量因素的实际关注度(True Reach);反映微友间谈话质量和传播速度因素的放大概率(Amplification Probability);反映用户的微博对网络粉丝影响因素的网络影响(Network Influence)。

Klout 指数已逐渐为推特和脸书用户接受,成为测试用户在推特和脸书内影响力的准官方指标,并推出若干 Klout 排行榜,如推特女士 TOP10 等。网上还流传提高 Klout 指数的秘籍技巧,指导网友如何提高影响力。诸如:尽量接触重要人物,远离草根人士;争取别人关注你,而不是去注意他人等这些五花八门的





图 10-9 网友 Cristen Perks 的 Klout 指数

内容。

值得一提的是 Klout 与推特和脸书的合作模式，是一种强强相容，优化般配。推特和脸书把精力放在扩大本身业务方面，而成员的影响力评价甚至引导成员如何提高影响力的业务，外包给合作伙伴 Klout 公司。

另外一家公司 AtImpress，专注于开发微博的应用平台，由 9th. be 推出"atimpress 爱影响"，目前尚在测试阶段。9th. be 曾经发表过一些和数据相关的微博应用产品，如"看看你的话痨指数""我的微博被转发几次"等。在这些微博附件的基础上，又推出了核心产品"AtImpress"，试图用数据量化每个人在社交网络内的影响力。

国内有关社交网络成员影响力的研究并不多见。2011 年 6 月，新浪微博推出"微数据"分析工具[9,10]，让成员对自己或周围的粉丝的影响力进行定量分析。按新浪的定义，个人的影响力是覆盖度、传播力、活跃度三者的综合体现，参见图 10-10。这项工作很有意义，增加了粉丝间的相互了解，迈出社交网络平台内影响力量化分析的可喜一步。如影星姚晨的影响力为 1309，新浪博客品牌的影响力为 844。



图 10-10　新浪微博"微数据"影响力分析工具

在新浪微博推出微数据之前，国内一家名叫微博风云的网站专门分析新浪微博用户影响力，据说使用人工智能的数据挖掘算法对此进行排名，使用的指标是：

(1) PR 值(People-Rank 值)是粉丝质量指数，PR>1 代表粉丝质量高于平均水平。

(2) 关注率：是指活跃用户关注的比例。如某微博关注率是 20%，代表 100 位活跃用户有 20 人关注该微博。详情请访问其网站。

这些微观分析，比较仔细，个性因素很强，反映了社交网络发展的方向，往往是单一平台的微分析，如对新浪微博的分析。作为个人或企业用户，有一定的实际意义，可以指导自己在网络内的行为和提高成效。问题是，目前在国内，很难有第三开发商像 Klout 进入推特和

脸书内，专门研究用户影响力指数这种商务模式。同时，新浪微博推出“微数据”分析工具后，由于自身数据等优势，AtImpress 和微博风云的市场空间也许会受到影响。正如 36 氪谈到：上述 3 个指数的不足之处在于仅支持新浪微博。希望早日能看到一个支持新浪微博、腾讯微博、人人网等诸多社交网站的中国社交影响力指数。

10.4.3　社交网络多平台影响力分析

社交网络成员在多平台间的影响力分析不同于 Klout 和新浪微博的微观分析，在美国、英国和日本等国开始引起关注，开始流行。星期日泰晤士报(Sunday Times)发表的大英社交网络的 2000 名社交排行榜(The Social List)就是一例，参见图 10-11。该排行榜基于网民在推特、脸书、LinkedIn 和 Foursquare 诸平台内的连接、推讯、共享、更新和聊天等行为进行统计，给出综合指标。星期日泰晤士报曾于 1989 年推出英国千人富豪榜，在全球影响较大。



图 10-11　英国社交网络的 2000 名社交排行榜(The Social List)网站截图

美国的一家网站 Famecount 综合脸书、推特和图片网库(YouTube)三家网络平台的排行指标，对国际名人和企业品牌列出唯一指数的排行榜。Famecount 在美国发行较为成功，但评价指标尚有以下问题。

(1) Famecount 指数是综合上述 3 项指标的加权平均数，没有反映成员信息传播的不均匀性；

(2) Famecount 主要目的是娱乐和商用，以致排行结果与原数据排次有一定偏差；

(3) Famecount 使用的图片网库(YouTube)指数，对于草根人士基本上没有意义；因为普通人，特别是中国人，很少把图片放到该平台上；

(4) Famecount 主要是面向西方社交网络，对中国相关的华语社交网络没有提及。

社交网络成员影响力-网熵指数排行系统(Social Wentropy Index Rank System)也是多平台的宏观分析，此项分析结果如英国的社交排行榜和美国的 Famecount，会引起公众特别是企业和政府的关注。就国内的情况看，新浪微博和博客没有一个统一的指标来评价其成

员对新浪整体的影响力。进一步来说，很难有人能把新浪、腾讯、搜狐、人人、百度等网络运营商召集在一起，研究一个统一的多平台的用户影响力指数，这就给社交网熵指数排行系统留下了市场空间。

该系统使用信息熵理论，科学地反映了成员对社交网络的影响力，明显比 Famecount 先进和实用。网熵指数排行系统主要面向中国和海外华人社交网络市场，填补了这项空白。社交网熵指数这一多平台的宏观分析系统与新浪、腾讯、搜狐、人人、百度等平台内部微观分析是一种共赢互补关系。希望今后能有机会结合 Klout 和新浪微博"微数据"等的方法，与这些运营商合作，直接使用诸家的统计数据，增加网熵指数排行精度，共同推出公正、科学、可持续性的中国社交网络成员影响力分析系统。

使用社交网熵指数排行的《美国社交网络金榜五人》一文在海外著名的华人网站文学城发表后，很快被选为推荐博客，参见图 10-12。2011 年 6 月 17～19 日（周五～周日）3 天内的访问量达到 4600 人次，列入本周文学城博客排行榜第 55 名。可见在海外华人的社交网络，此类排行也是空白，社交网熵指数排行系统颇受海外华人欢迎。

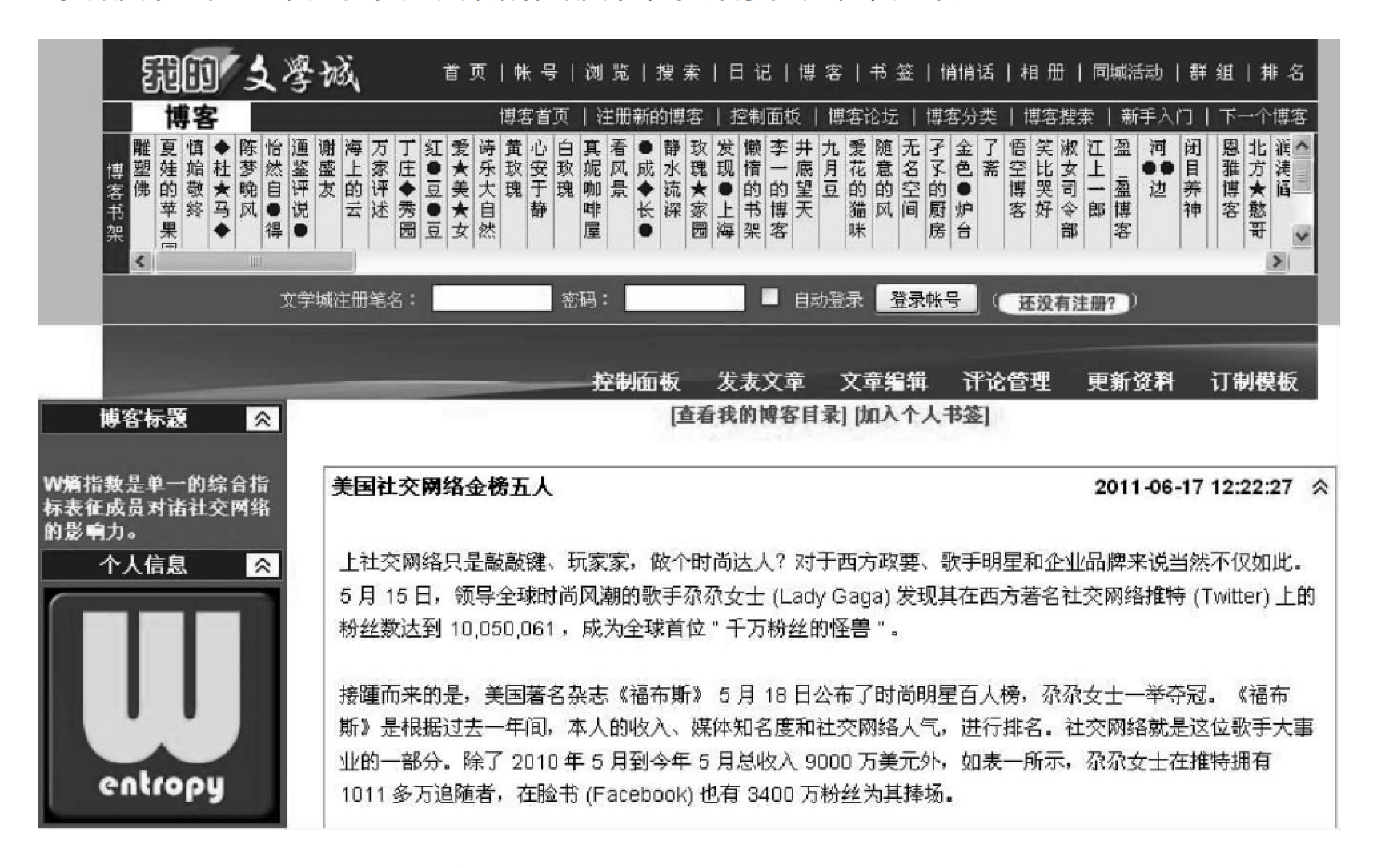


图 10-12　文学城博客登出社交网熵指数排行的美国社交网络金榜 5 人

2011 年 8 月 23 日，网熵科技在新浪博客、新浪微博、腾讯博客、腾讯微搏、搜狐博客和百度搜索上采集近 2000 组数据，根据这些社交网络成员的粉丝数、访问量和搜索文档数等指标，使用社交网熵指数的模型和系统，统计出这些成员的网熵指数，给出中国社交网络成员影响力两百人排行榜。介绍文章《中国社交网络影响力两百人》发表于《价值中国》上，该文由编辑推荐，列入当日该网站搏文排行榜。



金融大数据解决方案

11.1　金融信息化

金融信息化是构建在由通信网络、计算机、信息资源和人力资源 4 要素组成的国家信息基础框架之上，由具有统一技术标准，通过不同速率传送数据、语音、图形图像、视频影像的综合信息网络将具备智能交换和增值服务的多种以计算机为主的金融信息系统互连在一起，创造金融经营、管理、服务新模式的系统工程。

金融信息化行业的分类如下。

(1) 硬件，指实现数据的存储、处理、计算以及传输等，主要包括各类计算机主机设备、网络通信设备等。

(2) 软件，用来实现具体功能的各种计算机程序，即包括各类事务处理软件，如银行业务处理系统、工作流等；也包括诸如 ERP(企业资源管理)、CRM(客户关系管理)这类管理软件；以及用于决策支持的数据仓库、数据挖掘、统计分析等辅助分析的软件等。

(3) IT 应用服务，指综合运用软硬件技术，将其应用到业务运营、管理决策等领域中，以解决各类实际问题，提升效率及管理水平的手段和过程。

11.1.1　全球金融信息化发展历程

20 世纪 60 年代以来，银行、证券和保险行业纷纷开始用计算机代替手工作业，开启信息化历程。全球金融业信息化发展大体经历了以下 4 个阶段。

(1) 脱机业务处理。主要是实现银行业务的计算机辅助处理，其主要目的是节省查询时间和节约成本，这是金融企业信息化的起步阶段。

(2) 联机业务处理。信息技术在金融企业内部迅速渗透，主要是采用计算机网络技术，实现金融企业内部的联机业务处理，信息资源通过网络实现了共享。

(3) 经营决策信息化。即充分利用数据仓库等技术，实现了综合的客户信息分析，建立了电话银行、自助银行等新型服务体系，使基于 IT 的现代银行管理和业务体系趋于完善。

(4) 业务集成化和决策智能化。随着互联网和通信技术的飞速发展，国外银行纷纷开展了基于互联网技术的银行服务与产品创新，出现了网络银行、信用卡、ATM 卡、在线支付以及各种电子支票支付、网络保险、网上证券等新型产品和服务。

信息技术对人类经济社会的发展产生了重要影响，现代金融行业的发展更加离不开金融信息技术的支持。国外的金融信息化发展早已经进入业务集成和决策智能化阶段，经过信息技术的投资改造，欧美等国的银行业务发展能力极大提高，收益率增长明显。信息技术

给传统金融带来了新的活力。

11.1.2 我国金融信息化发展趋势

金融行业信息化的实质，是新兴的信息技术对传统金融业的一场经济革新，主旨在于把金融业变成典型的基于信息化技术的产业，信息系统成为金融产业战略决策、经营管理和业务操作的基本方式。我国金融信息化建设起步于20世纪80年代中期，经过三十多年的发展，目前已基本形成了比较完善的基于IT技术的金融服务体系。近几年来，中国金融业的经营环境发生了巨大变化，金融体制、经营理念、经营方式和管理模式发生了深刻变革，现代科学技术已经成为金融变革的主要推动力和支撑力。2013年，中国金融行业信息化投入达到511.6亿元，同比增长4.34%；2014年投资规模达到530亿元，同比增长3.60%。

近年来，我国银行业的整体实力和抗风险能力逐步增强，金融调控和监管不断加强。银行总资产及客户贷款数量稳步上升，银行总资产及贷款规模的增加既为银行带来了更多的收入和利润，又意味着更大的潜在风险。总的来说，银行业资产规模的扩大将会刺激银行信息化投入意愿的增加，其中一个重要原因就是风险敞口的增长。数据显示，2013年中国银行业信息化IT投入规模为371.5亿元，比2012年增长4.9%。银行业是金融行业IT投资比重最大的细分行业，而商业银行的IT投资在银行业中又占据主要部分，因此商业银行IT投入的稳定增长将影响着金融业IT投资的基本走势。近年来，银行信息化产业链保持着较高的景气度，发展空间较大。国家层面"电子化"和"网络化"的政策融合，也将促进银行业IT方面的大额投入。

目前，国内银行已初步建立起信息化平台，信息化基础设施建设框架已基本完成，国内银行业信息化顺利跨越了大规模基础设施建设的阶段，未来的银行IT信息化应用将向管理和服务型方向发展。银行将在风险管理、网络银行、金融审计和稽核、商业智能、决策支持等领域加快投入，这些领域在未来将成为银行业信息系统集成应用的重点市场。

数据一直是信息时代的象征。2011年5月麦肯锡全球研究院发布了报告《大数据：创新、竞争和生产力的下一个新领域》后，大数据的概念备受关注。金融业是大数据的重要产生者，交易、报价、业绩报告、消费者研究报告、官方统计数据公报、调查、新闻报道无一不是数据来源。金融业也高度依赖信息技术，是典型的数据驱动行业。互联网金融环境中，数据作为金融核心资产，撼动了传统客户关系、抵质押品在金融业务中的地位。例如，信用卡消费记录中早就包含消费时的位置信息，现在就可以被互联网金融利用。

与传统金融相比，大数据给互联网金融不仅带来了金融服务和产品创新，以及用户体验的变化，创造了新的业务处理和经营管理模式，对金融服务提供商的组织结构、数据需求与管理、用户特征、产品创新力来源、信用和风险特征等方面也产生了重大影响，显著提升了金融体系的多样性，也对金融监管和宏观调控等方面提出了新的课题。大数据的使用正在改变金融市场，也需要改变监管市场的方式，以保证市场参与者负责地使用大数据。例如，2010年5月的"闪电暴跌"令道·琼斯工业平均指数(Dow Jones Industrial Average)突然大跌，美国监管部门认为是高频交易造成了快速抛售引发的更多抛售。2013年4月23日的"无厘头暴跌"的缘由是美联社的Twitter账号发出巴拉克·奥巴马遭遇恐怖袭击的虚假消息：大数据中的一个数据点出错就能导致"无厘头暴跌"。

11.2　金融大数据综述

金融大数据是指集合海量非结构化数据,通过对其进行实时分析,可以为互联网金融机构提供客户全方位信息,通过分析和挖掘客户的交易和消费信息掌握客户的消费习惯,并准确预测客户行为,使金融机构和金融服务平台在营销和风控方面有的放矢。麦肯锡的研究显示,金融业在大数据价值潜力指数中排行第一。大数据决策模式对银行业更具针对性,发展模式转型、金融创新和管理升级等都需要充分利用大数据技术、践行大数据思维。

基于大数据的金融服务平台主要指拥有海量数据的电子商务企业开展的金融服务。大数据的关键是从大量数据中快速获取有用信息的能力,或者是从大数据资产中快速变现的能力,因此,大数据的信息处理往往以云计算为基础。目前,大数据服务平台的运营模式可以分为以阿里小额信贷为代表的平台模式和以京东、苏宁为代表的供应链金融模式。大数据的 4V 特点:Volume(大量)、Velocity(高速)、Variety(多样)、Veracity(精确)。

金融大数据模式广泛应用于电商平台,以对平台用户和供应商进行贷款融资,从中获得贷款利息以及流畅的供应链所带来的企业收益。随着金融大数据的完善,企业将更加注重用户个人的体验,进行个性化金融产品的设计。未来,金融大数据企业之间的竞争将存在于对数据的采集范围、数据真伪性的鉴别以及数据分析和个性化服务等方面。

目前,国内金融大数据领域发展较快的有阿里巴巴的金融电商平台、九次方的企业大数据交易、IBM 的 Watson 大数据人工智能等。而在金融大数据应用方面也是百花齐放,“激活数据,智创未来”是清数集团董事长赵勇博士提出的大数据口号,清数科技致力于研发金融大数据的应用,如协助某银行建立银行业务系统的统一数据分析平台,在 CRM、OA、门户网站、营销数据、信贷数据、交易数据、信用卡数据、呼叫中心数据等方面提供大数据智能分析挖掘服务。利用数据分析挖掘算法,对用户行为数据、用户群体分析等方面提供对应的商业分析,结合邮件、短信、线上精准推送等服务。

11.2.1　金融大数据的特征

(1) 网络化的呈现。在金融大数据时代,大量的金融产品和服务通过网络来展现,包括固定网络和移动网络。其中,移动网络将会逐渐成为金融大数据服务的一个主要通道。随着法律、监管政策的完善,随着大数据技术的不断发展,将会有更多、更加丰富的金融产品和服务通过网络呈现。支付结算、网贷、P2P、众筹融资、资产管理、现金管理、产品销售、金融咨询等都将主要通过网络实现,金融实体店将大量减少,其功能也将逐渐转型。

(2) 基于大数据的风险管理理念和工具。在金融大数据时代,风险管理理念和工具也将调整。例如,在风险管理理念上,财务分析(第一还款来源)、可抵押财产或其他保证(第二还款来源)重要性将有所降低。交易行为的真实性、信用的可信度通过数据的呈现方式将会更加重要,风险定价方式将会出现革命性变化。对客户的评价将是全方位、立体的、活生生的,而不再是一个抽象的、模糊的客户构图。基于数据挖掘的客户识别和分类将成为风险管理的主要手段,动态、实时的监测而非事后的回顾式评价将成为风险管理的常态性内容。

(3) 信息不对称性大大降低。在金融大数据时代,金融产品和服务的消费者和提供者之间信息不对称程度大大降低。对某项金融产品(服务)的支持和评价,消费者可实时获知该信息。

(4) 高效率性。金融大数据无疑是高效率的。许多流程和动作都是在线上发起和完成,有些动作是自动实现的。在合适的时间,合适的地点,把合适的产品以合适的方式提供给合适的消费者。同时,强大的数据分析能力可以将金融业务做到极高的效率,交易成本也会大幅降低。

(5) 金融企业服务边界扩大。首先,就单个金融企业而言,其最适合经营规模扩大了。由于效率提升,其经营成本必随之降低。金融企业的成本曲线形态也会发生变化。长期平均成本曲线,其底部会更快来临,也会更平坦更宽。其次,基于大数据技术,金融从业人员个体服务对象会更多。换言之,单个金融企业从业人员会有减少的趋势,或至少其市场人员有降低的趋势。

(6) 产品的可控性、可受性。通过网络化呈现的金融产品,对消费者而言,是可控、可受的。可控,是指在消费者看来,其风险是可控的。可受,是指在消费者看来,首先其收益(或成本)是可接受的;其次产品的流动性也是可接受的;最后消费者基于金融市场的数据信息,其产品也是可接受的。

(7) 普惠金融。金融大数据的高效率性及扩展的服务边界,使金融服务的对象和范围也大大扩展,金融服务也更接地气。例如,极小金额的理财服务、存款服务。支付结算服务等普通老百姓都可享受到,甚至极小金额的融资服务也会普遍发展起来。传统金融想也不敢想的金融深化在金融大数据时代完全实现。

11.2.2 金融大数据的机遇和挑战

1. 金融产业面对的机遇

金融业作为大数据的主要产生方,其中大数据可以创造的价值则不可估计:金融产业作为信息密集型的服务型产业,它所产生的交易记录、借贷记录、消费者信用记录报告等都是数据来源,而每一个从事金融产业的企业都会对自己的企业进行高规格的IT设施的投资,所以这些企业都会拥有较为庞大的数据信息库可以被利用。互联网的逐步普及,使得金融信息化的程度也在不断深化,电子银行、电子货币、快捷支付等金融产品和服务在迅速得到推广和扩散,金融产业的版图也不断再发生重组。在这种趋势下,必将会催生大量的金融数据,非结构化数据被纳入数据库,来自银行、电商、其他互联网金融公司的大量数据被收集,通过云计算对其进行整理和交互分析,产生多样化的用户数据结果。

从金融产业营销角度来说,大数据将能更清晰、数据化地得到客户的偏好和需求,通过定向营销或个性化推荐吸引客户、增加客户黏性。而从风险管理角度来说,通过多渠道、多角度的数据来源和对交易数据的深度挖掘,金融业将能够做到实时监控,及时排查潜在金融风险,降低风险管理成本,提高监管效率。

大数据时代的来临必将是金融业发展的绝好机遇。

2. 金融产业面对的挑战

在面对绝佳的机遇的同时,也会面临诸多的挑战。

(1) 大数据的兴起会对传统金融机构形成压力。当前的环境下,客观上来看,已经降低了金融服务业的准入门槛,传统意义上的非金融机构更多的是想利用自身的优势在金融市场中占得一席之地。相反,传统金融机构被困于已有的组织架构和陈旧规则,而不能发掘自己的价值与潜力,在金融竞争的浪潮中处于劣势。例如,支付宝已经在网络购物支付领域处

于领头羊的绝对优势地位。互联网金融企业在大数据时代，可以获得更多非结构化数据，不再限于用户的现金流水等结构化数据，用户的互联网行为都被收集在大数据库中，对用户的分析将更加完善和真实化。同时凭借互联网的快速扩展，便捷的线上互联网金融产品对传统金融机构产业产生了巨大冲击。

（2）数据基础设施的挑战。以一个完整的支付链条为例，一笔支付业务可以分为交易前、交易、清算、结算和交易后 5 个阶段，进一步可以将上述环节细分为十多个环节，每一个环节基本上都有独立的机构进行经营管理，同时每一个环节也必将产生大量的数据往来，而这些大量的数据往来必然会冲击各个环节上的机构的基础数据设施。

（3）金融数据的安全性日益突出。网络大数据为金融业务的发展提供了便利，同时也为金融犯罪降低了成本。大量的数据收集整理同时也意味着更大的数据泄漏风险。虽然诸多金融机构都致力于在数据安全方面上进行大量投资，但是无奈金融业务的链条较长，各环节一旦有微小瑕疵都会造成金融财产的不安全。早在 2010 年，中国香港八达通公司对其拥有的进两百万的客户数据信息私下销售，引起了港民的极大不满，也暴露了金融数据的安全隐患。

3. 金融产业的应变对策

应对大数据对金融产业的冲击的对策的关键就在于，要坚定信念地发挥传统金融机构的优势，同时要寻求新型金融机构的创新之路，做到多角度全方位的发展。从顶层设计入手，面向全局考虑发展，并且要时刻保持以客户需求为导向，积极努力地去构造金融机构自身的大数据规模，同时也要做到保障信息安全。而这个规模要从两大类设备实施：一是软基础设施，主要包含从事金融产业和大数据处理的大量的人力资源，要保证这些资源有足够的储备，以此来保证金融机构有足够多的智力、技术资本来保证金融大数据的原汁原味，并且需要加强内部控制，保证金融机构所掌握的用户数据不被泄漏；二是硬基础设施，主要包含基础 IT 设备和信息安全防范系统。运用基础 IT 设备来集约化地完成金融机构内外的金融数据的收集、汇总、处理及分析，使机构、用户双方都可以以最快的、最便利的方式调用自己本身需要的信息。而完备的信息安全防范系统将通过强化身份认证、数字证书等安全认证，加快信息安全等级保护制度的建立，切实保护数据安全。

随着国内网购市场的迅速发展，淘宝网等众多网购网站的市场争夺战也进入白热化状态，网络购物网站也开始推出越来越多的特色产品和服务。以余额宝为代表的互联网金融产品在 2013 年刮起一股旋风，截至目前，规模超 1000 亿元，用户近三千万，相比普通的货币基金，余额宝鲜明的特色当属大数据。以基金的申购、赎回预测为例，基于淘宝和支付宝的数据平台，可以及时把握申购、赎回变动信息。另外，利用历史数据的积累可把握客户的行为规律。淘宝网的“阿里小贷”更是得益于大数据，它依托阿里巴巴（B2B）、淘宝、支付宝等平台数据，不仅可有效识别和分散风险，提供更有针对性、多样化的服务，而且批量化、流水化的作业使得交易成本大幅下降。每天，海量的交易和数据在阿里的平台上跑着，阿里通过对商户最近 100 天的数据分析，就能知道哪些商户可能存在资金问题，此时的阿里贷款平台就有可能出马，同潜在的贷款对象进行沟通。

11.3　金融大数据平台总体架构

金融大数据服务平台分为数据应用、数据计算、数据管理、数据源 4 个层面，如图 11-1 所示。数据应用、数据管理层需要整合和兼容原有系统进行延续和提升；数据计算层是需

要新开发的内容，在实时数据分析应用与历史数据挖掘方面具有潜在的研究方向。

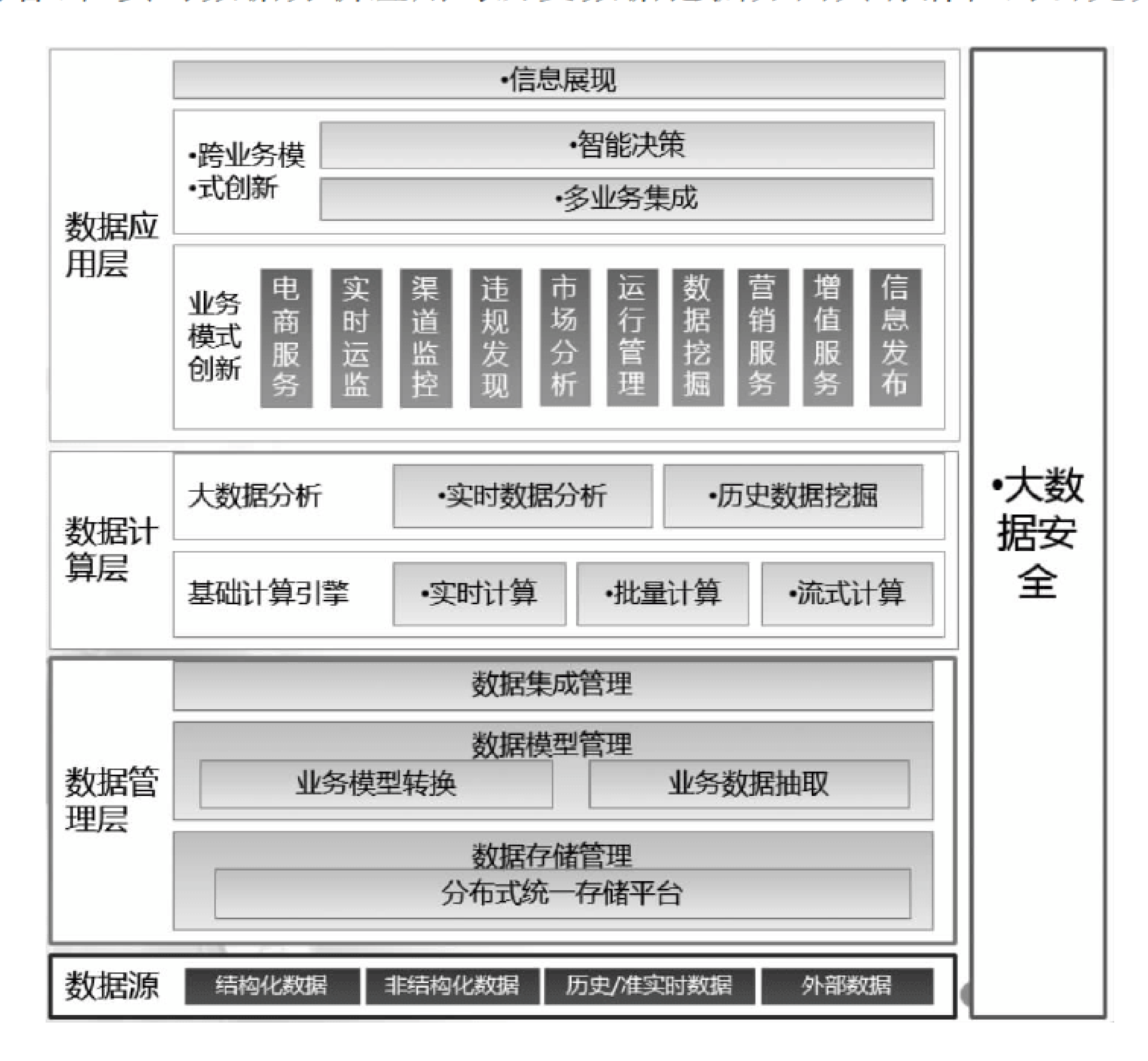


图 11-1　金融大数据服务平台

11.3.1　建设原则和目标

平台是大数据的基础实施，其建设、设计和系统实现过程中，应遵循如下指导原则。

(1) 经济性：基于现有场景分析，对一定时间段内的数据量进行合理评估，确定大数据平台规模，后续根据实际情况再逐步优化扩容。

(2) 可扩展性：架构设计与功能划分模块化，考虑各接口的开放性、可扩展性，便于系统的快速扩展与维护，便于第三方系统的快速接入。

(3) 可靠性：系统采用的系统结构、技术措施、开发手段都应建立在已经相当成熟的应用基础上，在技术服务和维护响应上同用户积极配合，确保系统的可靠；对数据指标要保证完整性，准确性。

(4) 安全性：针对系统级、应用级、网络级，均提供合理的安全手段和措施，为系统提供全方位的安全实施方案，确保企业内部信息的安全。大数据技术必须自主可控。

(5) 先进性：涵盖结构化、半结构化和非结构化数据存储和分析的特点。借鉴互联网大数据存储及分析的实践，使平台具有良好的先进性和弹性。支撑当前及未来数据应用需求，引入对应大数据相关技术。

(6) 平台性：归纳整理大数据需求，形成统一的大数据存储服务和大数据分析服务。利用多租户，实现计算负荷和数据访问负荷隔离，多集群统一管理。

(7) 分层解耦：大数据平台提供开放的、标准的接口，实现与各应用产品的无缝对接。

金融大数据平台，通过采集银行内部与外部、静态与动态的各类金融数据，搭建适于大数据存储与分析的 Hadoop 集群，对金融数据采取合适的预处理方式，利用数据挖掘技术得

出隐藏在海量数据后的、有价值的潜在规律，以丰富的可视化模型向客户进行展现，在此基础上实现精准营销、统一广告发布、业务体验优化、客户综合管理、风险控制等金融业务应用。由此，提升金融业务的水平和效率，推进银行业务创新，降低银行管理和运行成本。具体技术目标包括以下几个方面。

(1) 构建金融数据采集工具：大数据分析需要收集来自银行内部的和外部的、静态的和动态的各种金融数据，为此构建各类金融数据采集工具，如动态采集 SDK、日志提取分析工具、外部数据导入工具等。

(2) 搭建 Hadoop 大数据集群：搭建 Hadoop 大数据集群，是建设“金融大数据平台”的基础。利用多台性能较为一般的服务器，组成一套基于 HDFS 和 Map-Reduce 机制的集群，并根据需要在其上安装 Hive、HBase、Sqoop、Zookeeper 等软件。

(3) 实现分析挖掘算法：支持 Hadoop 的分析挖掘算法，是“金融大数据平台”的一个关键组成部分。在利用传统数据挖掘技术的基础上，实现包括抽象的数学算法(如关联算法、分类算法、聚类算法、时序分析算法等)，以及在此基础上针对金融业务的专业算法(如客户行为特征模型、效果分析模型等)，作为进一步构建抽象模型和金融专业模型的基础。

(4) 构建分析挖掘模型：支持 Hadoop 的分析挖掘模型，是“金融大数据平台”的另一关键组成部分。在上一步基础上，快速构建抽象的数学模型(如神经网络模型、事物关联模型等)，以及针对金融业务的专业模型(如精准营销模型、广告效果评估模型等)。

(5) 构建 ETL 工具：数据预处理也是“金融大数据平台”需要解决的问题之一。利用市场上已有的数据预处理成果，构建一个支持 Hadoop 的 ETL 工具，实现包括规范化、数据抽样、数据排序、汇总、指定因变量、属性变换、数据替换、数据降维、数据集拆分、离散化等功能。

(6) 实现可视化展现工具：“金融大数据平台”上的分析结果将主要采用丰富多彩的可视化形式向用户进行可视化展现。可以支持：分类树图、视觉聚类图、关联图、序列图、回归图等多种可视化形式。

(7) 实现金融业务应用：将分析挖掘的结果集成到具体的银行业务系统中，如精准营销系统、统一广告发布平台、业务体验优化系统、客户综合管理系统、风险控制系统等。具体方式既可以是实现某个独立的新业务系统，也可以是在现有系统中实现一个或多个新模块，从而扩充或提升原有的功能。

11.3.2 金融大数据业务架构

“金融大数据平台”由数据采集层、数据存储层、分析挖掘层和业务应用层组成，总体框架如图 11-2 所示。

(1) 数据采集层：负责从各类数据源中提取、导入数据，主要产品包括：动态采集 SDK、日志提取分析工具、外部数据导入工具、其他数据提取工具等。

(2) 数据存储层：负责将预处理后的数据进行存储，主要由可进行横向扩展的 Hadoop 集群构成，另外辅之以关系数据库做数据中转、元数据存储、供某些软件使用等用途。

(3) 分析挖掘层：负责金融数据的建模、挖掘、评估和发布，核心是实现两类数据挖掘的算法和模型：一类是抽象的数学算法及模型，另一类是在此基础上针对金融业务的专业算法和模型。

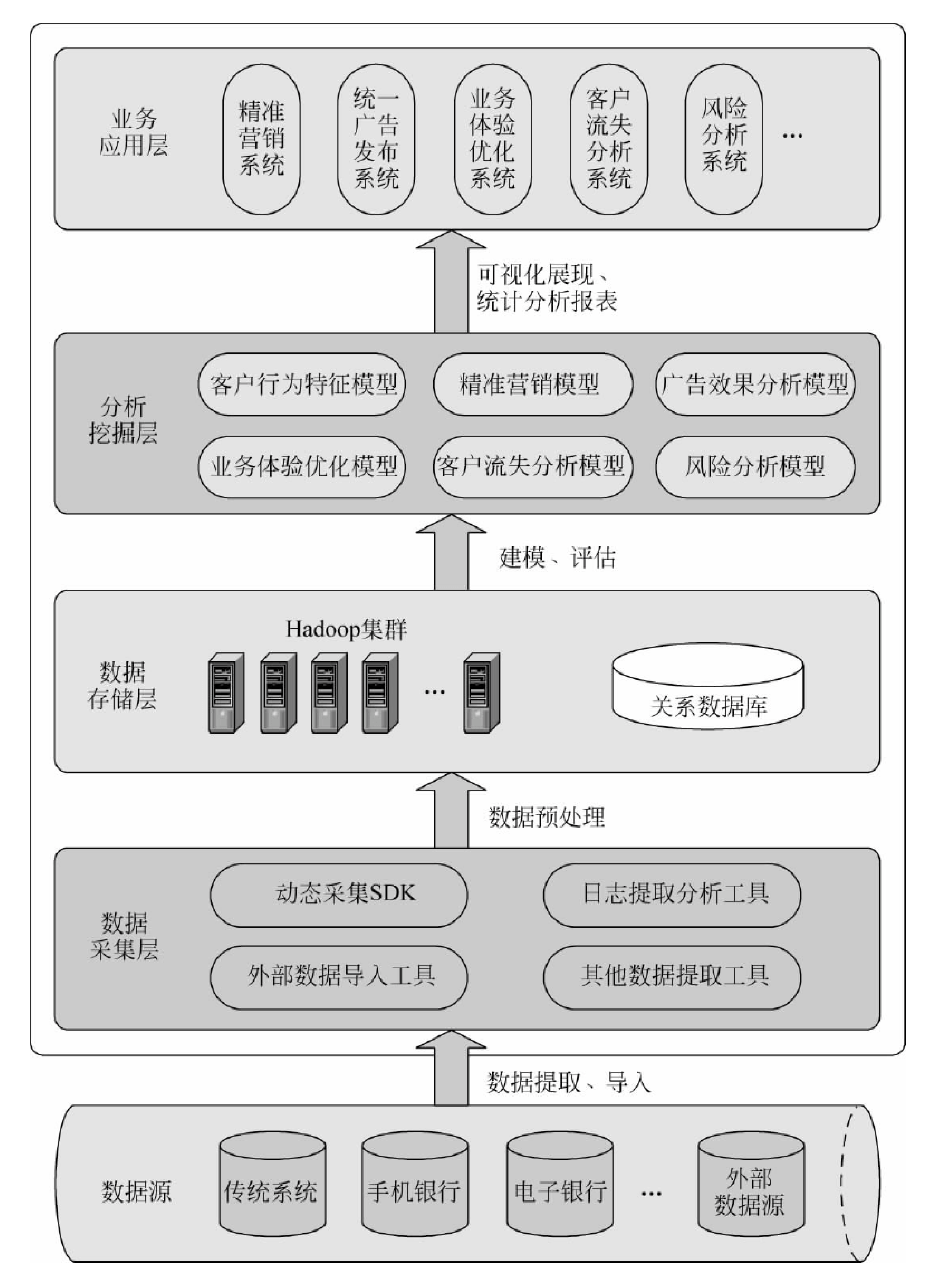


图 11-2　金融大数据总体框架

（4）业务应用层：负责将分析挖掘结果的可视化展现形式，集成到相应的金融业务系统中。

另外，在数据采集层和数据存储层之间，由 ETL 工具负责数据预处理任务；在分析挖掘层和业务应用层之间，由可视化展现工具负责分析挖掘结果的可视化展现任务。

11.3.3　金融大数据技术架构

“金融大数据平台”的技术架构采用多层次形式，如图 11-3 所示。

数据源包括各类动态数据（如行为数据）、静态数据（如属性数据）、日志文件以及其他数据等，可以是结构化的、半结构化的和非结构化的数据。

在数据采集层，各采集工具根据具体情况采用不同的技术实现方式，如对动态数据的采

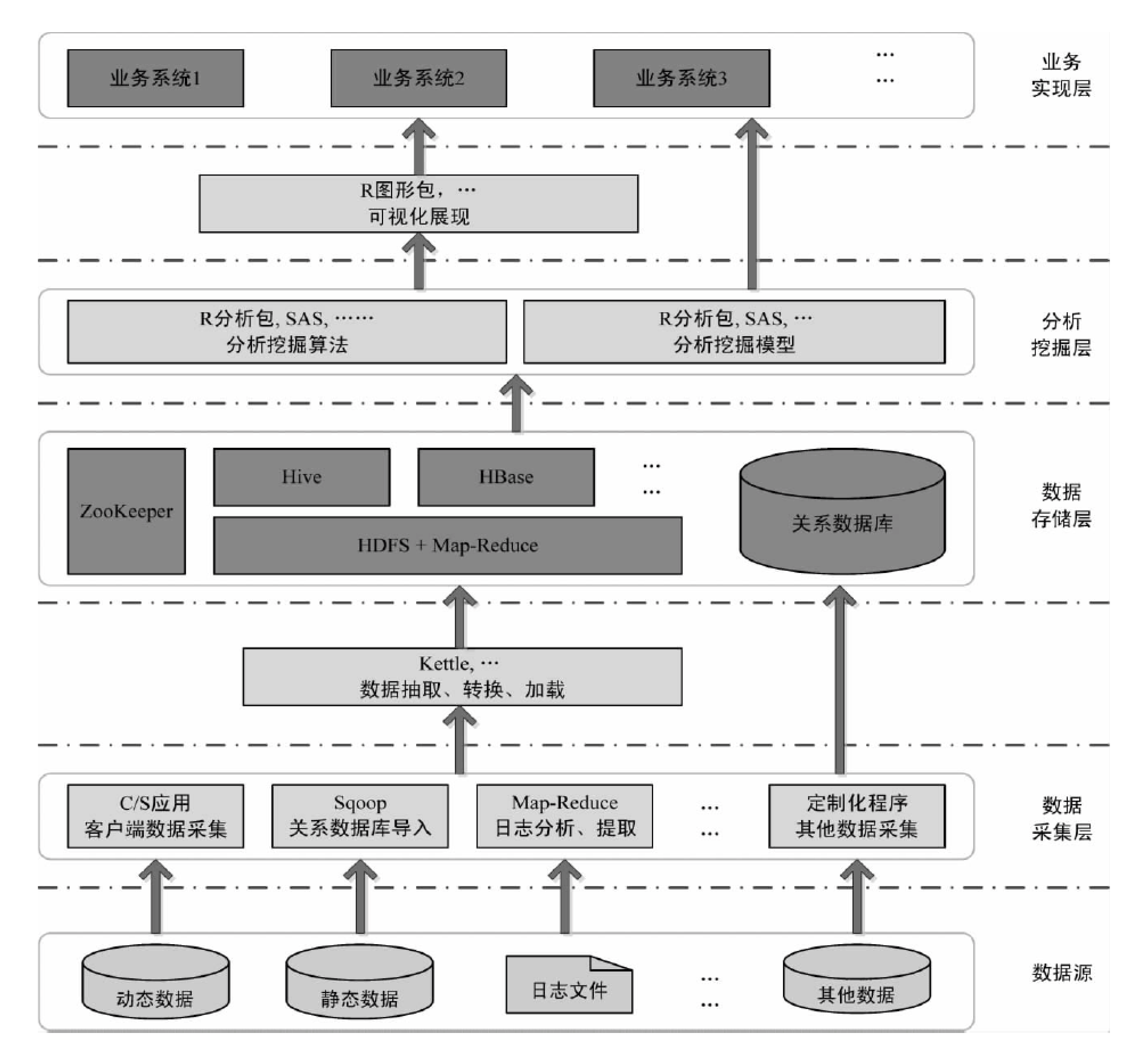


图 11-3 金融大数据技术架构

集，使用 C/S 架构的客户端采集 SDK，对日志文件使用 Map-Reduce 方式的分析提取工具，对静态数据按 Sqoop 方式从关系数据导入，对其他数据则使用定制化程序，等等。

ETL（数据抽取、转换、加载）将采集到的各种数据整合成统一的数据模型，包括数据清洗、数据转换、数据规约、数据集成等。为加快项目进度和保证项目质量，初步决定在某个支持 Hadoop 的开源 ETL 产品（如 Kettle）的基础上进行二次开发。

在数据存储层，Hadoop 集群使用 Hadoop 技术生态圈的诸多关键技术，包括：分布式存储 HDFS 系统、并行处理 Map-Reduce 机制、NoSQL 数据库 HBase、数据仓库 Hive、协调系统 Zookeeper 等。此外，还需用到关系数据库担任数据中转、元数据存储、供某些软件使用等用途。

分析挖掘层的任务是在 Hadoop 集群实现各种分析挖掘算法和分析挖掘模型。算法和模型有两类，一类是抽象的数学算法（如聚类算法、关联分析算法）和数学模型（如神经网络模型、事物关联模型等），另一类是在此基础上构建的专业算法（如金融客户分类算法、效果评估算法）和专业模型（如客户行为特征模型、效果评估模型）。为加快项目进度、保证项目质量和扩大适应范围，初步决定在 SAS 和 R 的分析挖掘包的基础上实现算法接口，并利用算法接口构建大部分模型，其余部分视实际情况而以自主研发方式构建。

可视化展现将分析挖掘结果面向用户进行各种可视化展现(如散点图、直方图、分布图、饼图等),分析挖掘的质量也决定着展现的质量。为加快项目进度,初步决定在某个可视化展现开源产品(如 R 的图形包)的基础上进行二次开发。

在业务实现层,分析挖掘结果集成到相应的金融业务系统中。具体方式既可以是实现某个独立的新业务系统,也可以是在现有系统中实现一个或多个新模块,从而扩充或提升原有的功能。

11.3.4 金融大数据网络架构

"金融大数据平台"采用集中部署方式,硬件环境由 Hadoop 集群服务器和数据库集群组成,如图 11-4 所示。

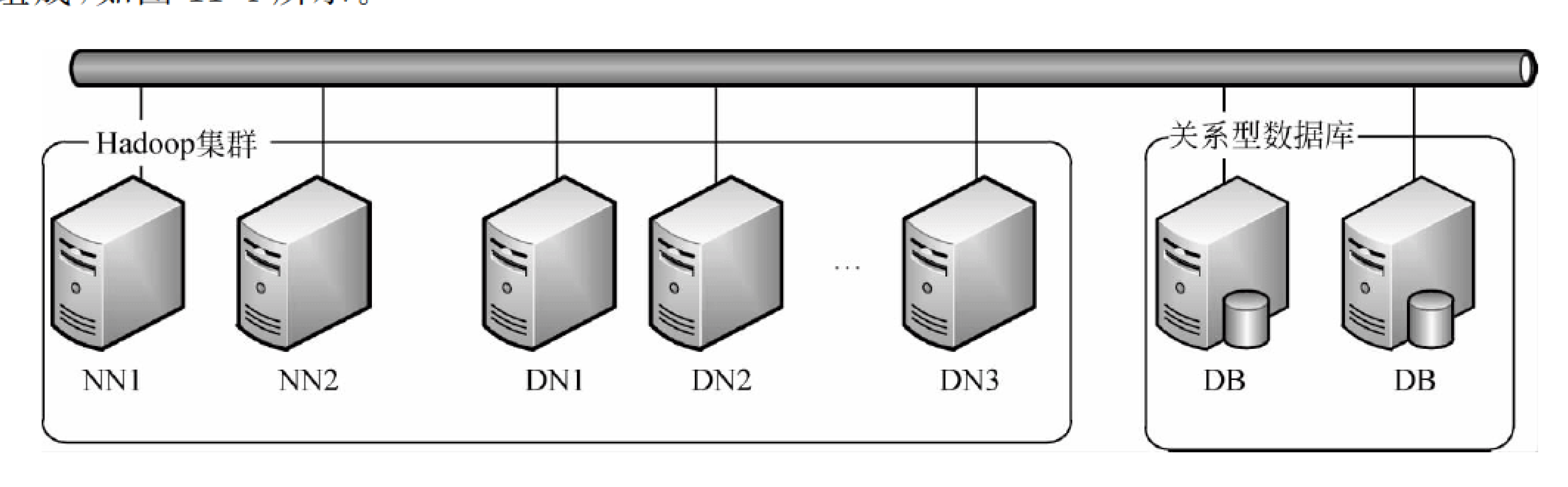


图 11-4 金融大数据物理架构

其中,Hadoop 集群包括两个 NameNode(主从方式)和多个 DataNode(最少三个,以后根据需要增加);NameNode 用于管理数据在 DataNode 上的分配,而 DataNode 用于数据的存储。NameNode 和 DataNode 采用相同的配置,运营环境中建议为:CPU 为两块×16 核,主频 2~2.5GHz,内存 128GB,硬盘 12 块×2TB。

数据库集群包括两台数据库服务器,采用双机热备方式。其配置建议为:CPU 为两块×16 核,主频 2~2.5GHz,内存 64GB,硬盘 12 块×2TB。

11.4 金融大数据分析

11.4.1 银行风险管理状况分析

银行面临的风险是风险管理工作的对象,因此,在进行风险管理工作之前,深刻理解面临的风险状况是必不可少的。随着全球经济进入下行周期以及银行市场化经营的不断深入,我国银行面临的风险也在不断加剧。不仅信用风险形势严峻,而且市场风险和操作风险也呈现更复杂的趋势。本节通过对银行面临的风险状况的分析,有助于更好地理解在《办法》推进实施背景下,全面提高风险管理水平的重要意义。

1. 信用风险

信用风险是指债务人或交易对手未能执行合同所规定的义务或信用质量发生变化,从而给债务人或金融产品拥有人造成经济损失的风险。银行作为信用中介机构,信用风险一直都是其所面临的最主要风险。目前,我国银行面临着严峻的信用风险,主要表现特征如下。

首先，信贷集中度过高，中长期贷款比重愈来愈大，信贷资金投向过于集中且行业重叠。在行业上，主要投向了铁路、公路和机场以及地方政府融资平台构造的基础设施行业、房地产业；在地域上，则较多投向了沿海经济发达地区。而银行贷款集中度过高，行业或地区出现周期性衰退就将导致大量信贷资金无法收回，这也就在一定程度上增加了银行的信用风险。其次，我国银行信用风险已进入爆发周期。国内外经验表明，信用风险 30 万亿的金额已经开始蔓延甚至侵入银行体系，形成极大危害。还有我国银行业存贷款期限错配严重。因为中长期贷款具有较高的信用风险，所以存贷款期限不仅可以作为反映流动性风险的重要指标，还可以在一定程度上作为反映信用风险的指标。从理想的状态考虑，短期贷款应与短期存款匹配，而中长期贷款应与长期存款匹配，但是在我国，银行存贷款期限错配趋势明显，中长期贷款占定期存款的比例已由 2003 年的 69%上升到 99%。

2. 市场风险

在我国，银行被禁止投资股票、期货等金融领域，因此我国银行面临的主要市场风险是利率风险和汇率风险。随着我国利率、汇率管理制度的逐步废除，市场化利率、汇率制度的逐渐形成，银行的利率自主权不断扩大，利率和汇率风险将成为我国银行未来面临的主要风险之一。

长期以来，我国实行严格的利率管制，银行存款相对稳定，支付能力一般不会出现问题。2012 年 6 月中国人民银行将利率的上下空间已经打开，利率市场化进程已经完成近 70%，银行间市场利率波动性不断加大，利率波动性上升，必然导致银行间竞争加剧，资金流动更加频繁，存款稳定性大幅度下降，在我国尚未建立起完备的存款保险制度的情况下，对银行的流动性提出了严峻的考验；当前，我国金融市场还不发达，资金来源和运用渠道单一，银行短时间内调整资产负债结构的能力有限，同时又缺乏对利率风险的保值工具和手段。因此，在利率波动加大后，银行将面临较大的利率风险。目前，我国的汇率制度正在从单一的有管理的汇率制度向市场化的浮动汇率制度转变，市场化的浮动汇率制度会使汇率波动的范围增大，我国银行面临的汇率风险也会因此而加大。

3. 操作风险

因人员、系统、流程和外部事件所引发的风险，根据《办法》定义均属于操作风险范畴。操作风险具有普遍性和非盈利性特征，它存在于银行业务的各个环节，操作风险的产出并不能为银行带来盈利，但是在业务办理过程中，银行又不可避免会发生并承担相应的损失。在当前经济下行及经营环境竞争加剧的背景下，银行违规操作导致重大案件发生的压力有增无减，未来操作风险形势将十分严峻。

从经济周期规律来看，银行操作风险及案件多发与实体经济不景气之间有着正相关关系，一些在经济高速发展时期被掩盖、被忽视的银行风险，很可能随着经济进入下行区间而水落石出，银行内控失效诱发的员工操作风险可能集中暴露，一些银行客户可能因为经济困难、资金链断裂而骗贷跑路。随着社会资金紧张，银行员工卷入民间借贷、非法集资等风险也会上升。从银行经营环境来看，人民银行在 2012 年 6 月和 7 月先后两次降息并扩大贷款利率浮动范围，标志着利率市场化改革步伐加快，存贷款利差收窄趋势明显，同时，随着金融改革向纵深推进，直接融资市场对银行优质客户的分流效应不断显现，银行市场竞争将更加激烈，银行机构员工规避监管、违规操作的外部驱动力增加，操作风险加大。

《商业银行资本管理办法》的实施对银行资本管理提出了更为严格、明确的要求。其中，首次将操作风险纳入我国银行业的资本监管框架，对操作风险的资本占比制定了具体标准，除第一支柱要求对操作风险计提资本外，在第二支柱中，监管部门还可以更加对单家机构的操作风险管理水平和操作风险事件的发生情况，提高其监管资本要求，这表明，银行的操作风险管理水平、案件防控情况将直接决定其资本消耗，进而影响银行各项业务的拓展能力。因此，各银行有必要将操作风险防控放在更加突出的位置。

4. 流动性风险

当银行出现流动性不足时，在极端情况下会导致银行资不抵债而破产清算。2008 年爆发的国际金融危机即是流动性风险爆发的突发性和银行业流动性管理的粗放性的集中体现。银行流动性风险按发生的原因包括由资产业务引起的流动性风险和由负债业务引起的流动性风险两种。我国银行业流动性风险出现一般有两种情况：一是银行确实没有足够的资金来满足存款人的日常取款需要；另一种情况是银行的资产治理不善，银行一时没有足够的能力将投放到其他项目中的资金调过来，暂时出现了流动性的困难。就目前我国银行业的情况来看，资产方面主要存在有短期贷款比例较低、中长期较高的现象，这种资金来源和运用期限出现了严重的错配为引发流动性风险带来隐患。《商业银行资本管理办法》引入了巴塞尔 III 中的流动性覆盖率和净稳定融资比率两个新监管指标，这将对银行的成本控制、盈利能力以及金融市场的流动性都会产生直接影响。因此，国内银行应高度警惕这种由于资本结构造成的现金流不足问题，加强流动性管理。

5. 其他风险

目前，无论从宏观上还是从微观上，我国银行体系已经累积了很大的风险，随着金融市场化和金融全球化的不断发展，这种风险的压力还将继续增加。尽管长期以来，无论银行不良资产如何巨大，银行风险并未对我国的经济发展带来特别严重的实质性影响，但中国式的银行危机已让国内外学者和业界忧心忡忡，主要原因在于以下几点：一是持续高速的经济增长及国家究底风险等制度性因素极大地掩盖了银行体系的风险，一是爆发必将对银行体系甚至经济实体带来无法估计的伤害。多年来，由于经济持续高速增长极大地扩张了社会总财富，我国居民高储蓄偏好更使得银行存款增长速度高于经济增长速度，银行业在不良贷款不断累积下仍能正常运转，使风险始终处于潜在状态；同时，我国银行风险实质上最终由国家承担，人民币在资本项下不能自由兑换，利率尚未完全市场化，存款保险制度未建立，银行退出机制缺乏等，也大大掩盖了银行风险。二是非银行体系的“影子银行”严重冲击中国的银行业，以民间借贷、地下金融、理财产品等近三十万亿的金额已经开始蔓延甚至侵入银行体系，形成极大危害。三是自 2012 年下半年以来，以云南、四川、上海等地区为代表出现地方政府债券大面积违约现象，利息支付不出。

11.4.2 金融大数据风险管理云平台

金融大数据风险管理云平台有利于提高金融机构稳定收益、有效控制风险，并具有快速决策和解决问题的能力，提升整体工作效率，改善流程，降低运营成本。从技术上讲，金融风险管理云平台就是利用云计算和大数据系统模型，将金融机构的数据中心与客户端分散到云里，从而达到提高自身系统运算能力、数据处理能力，改善客户体验评价，降低运营成本的

目的，如图 11-5 所示。

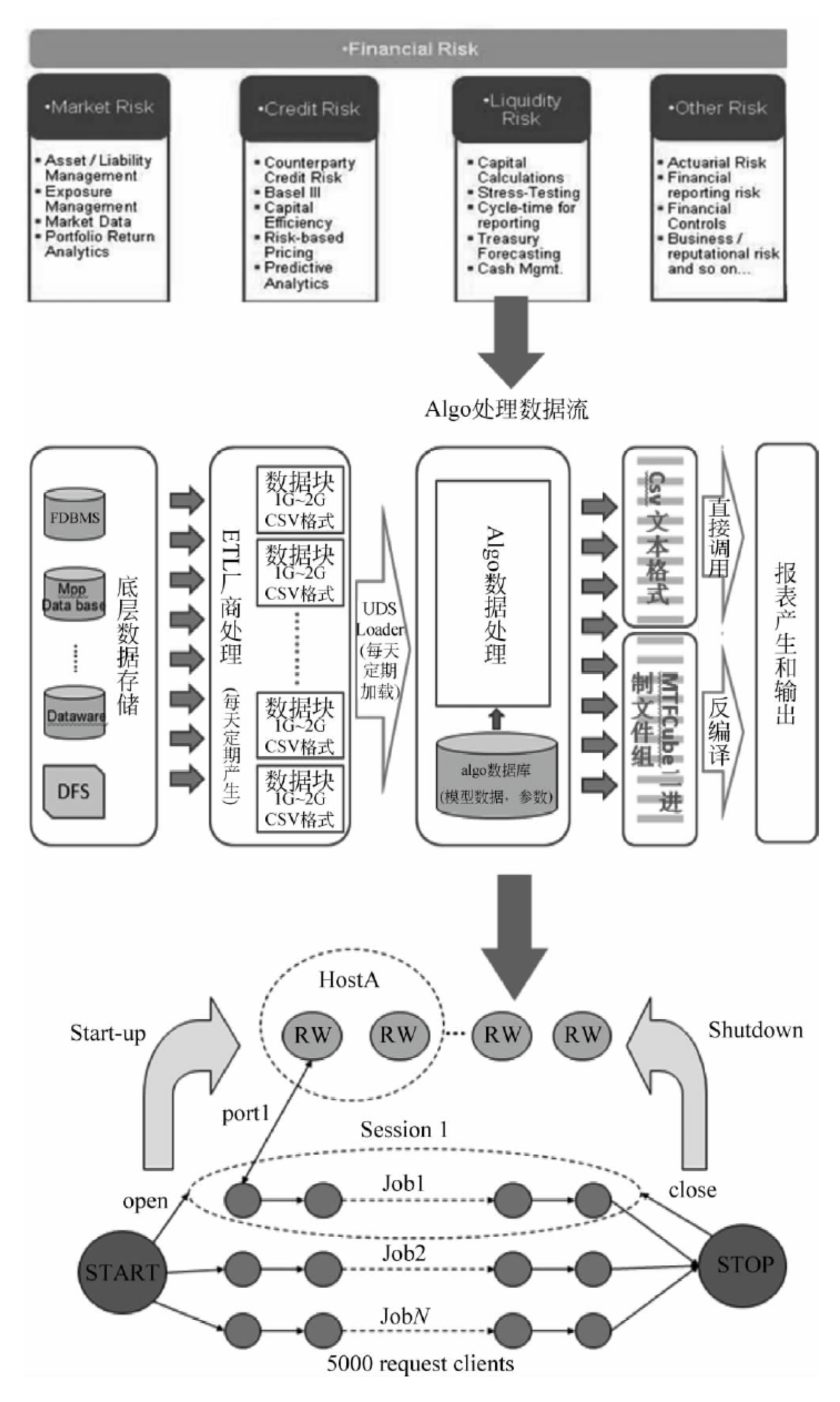


图 11-5　金融风险管理大数据云平台

1. 金融数据处理云应用

（1）构建云金融信息处理系统，降低金融机构运营成本。云概念最早的应用便是亚马逊（Amazon）于 2006 年推出的弹性云计算（Elastic Computer Cloud ES2）服务。其核心便是分享系统内部的运算、数据资源，以达到使中小企业以更小的成本获得更加理想的数据分析、处理、储存的效果。而网络金融机构运营的核心之一，便是最大化地减少物理成本和费用，提高线上（虚拟化）的业务收入。云计算可以帮助金融机构构建“云金融信息处理系统”，

减少金融机构在诸如服务器等硬件设备上的资金投入,使效益最大化。

(2) 构建云金融信息处理系统,使不同类型的金融机构分享金融全网信息。金融机构构建云化的金融信息共享、处理及分析系统,可以使其扩展、推广到多种金融服务领域。诸如证券、保险及信托公司均可以作为云金融信息处理系统的组成部分,在全金融系统内分享各自的信息资源。

(3) 构建云金融信息处理系统,统一网络接口规则。目前,国内金融机构的网络接口标准大相径庭。通过构建云金融信息处理系统,可以统一接口类型,最大化地简化诸如跨行业务办理等技术处理的难度,同时也可减少全行业硬件系统构建的重复投资。

(4) 构建云金融信息处理系统,增加金融机构业务种类和收入来源。上述的信息共享和接口统一,均可以对资源的使用方收取相关的费用,使云金融信息处理系统成为一项针对金融系统同业企业的产品,为金融机构创造额外的经济收入来源。

2. 金融机构安全系统的云应用

基于云技术的网络安全系统也是云概念最早的应用领域之一。现如今,瑞星、卡巴斯基、江民、金山等网络及计算机安全软件全部推出了云安全解决方案。其中,占有率不断提升的 360 安全卫士,更是将免费的云安全服务作为一面旗帜,成为其产品竞争力的核心。

所以说,将云概念引入到金融网络安全系统的设计当中,借鉴云安全在网络、计算机安全领域成功应用的经验,构建"云金融安全系统"具有极高的可行性和应用价值。这在一定程度上,能够进一步保障国内金融系统的信息安全。

3. 金融机构产品服务体系的云应用

通过云化的金融理念和金融机构的线上优势,可以构建全方位的客户产品服务体系。例如,地处 A 省的服务器、B 市的风险控制中心、C 市的客服中心等机构,共同组成了金融机构的产品服务体系,为不同地理位置的不同客户提供同样细致周到的产品体验。这就是"云金融服务"。

事实上,基于云金融思想的产品服务模式已经在传统银行和其网上银行的服务中得到初步的应用。金融机构可通过对云概念更加深入的理解,提供更加云化的产品服务,提高自身的市场竞争力。

例如,虽然各家传统银行的网上银行都能针对客户提供诸如储蓄、支付、理财、保险等多种不同的金融服务,但作为客户,其同一种业务可能需要分别在多家不同的银行平台同时办理。当有相应的需求时,就需要分别登录不同的网上银行平台进行相关操作,极其烦琐。而云金融信息系统,可以协同多家银行为客户提供云化的资产管理服务,包括查询多家银行账户的余额总额、同时使用多家银行的现金余额进行协同支付等,均可在金融机构单一的平台得以实现。如此一来,将会为客户提供前所未有的便利性和产品体验。

11.4.3 大数据征信

1. 大数据征信特征

大数据征信体系如图 11-6 所示。大数据征信体系具有覆盖面广、信息维度丰富、数据获取实时动态的优势。个人信贷风险评估主要从身份识别、还款意愿、还款能力三方面进行评估,大数据征信相对于传统线下的采集和整合更加全面和准确,其信用评估结果更加科

学，大数据征信与传统征信相比具有以下三方面的优势。

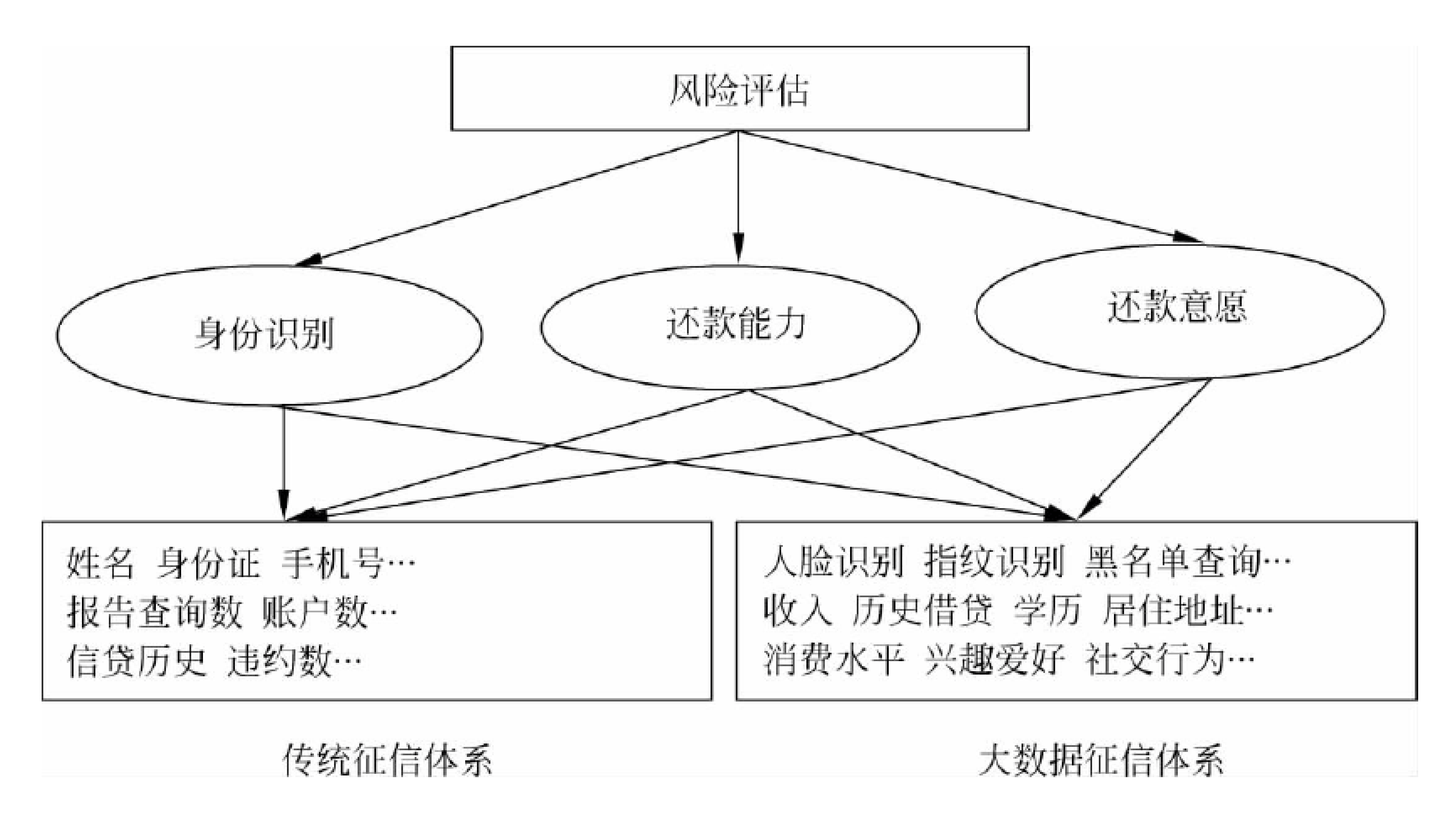


图 11-6　大数据征信体系比较图

（1）数据主要来源于互联网，互联网覆盖人群广泛，通过互联网获取数据，弥补了传统征信体系的不足，能够有效拓展业务。

（2）丰富了数据维度和种类，传统征信数据主要采集身份信息、信贷信息、非金融负债信息三类，以及部分公共信息，在大数据征信系统中，信用评估的来源更加广泛，社交网络与电子商务行为中产生的海量数据，都能给用户行为提供侧面支持。

（3）大数据挖掘获得的数据具有实时性、动态性，能够实时监测到信用主体的信用变化，企业可以及时拿出解决方案，避免不必要的风险。

大数据征信评估个人信用注重强相关信息，忽略弱相关信息。通过大数据技术手段可以挖掘申请人多维度信息，包括姓名、性别、年龄、电话、身份证件、家庭住址、职业、学历、信贷记录、支出、消费偏好、兴趣爱好、社交行为等信息。并不是所有数据都对个人信用评估有参考价值，数据采集的越多，审核纬度越多，个人信用评估模型越失真，如图 11-7 所示。

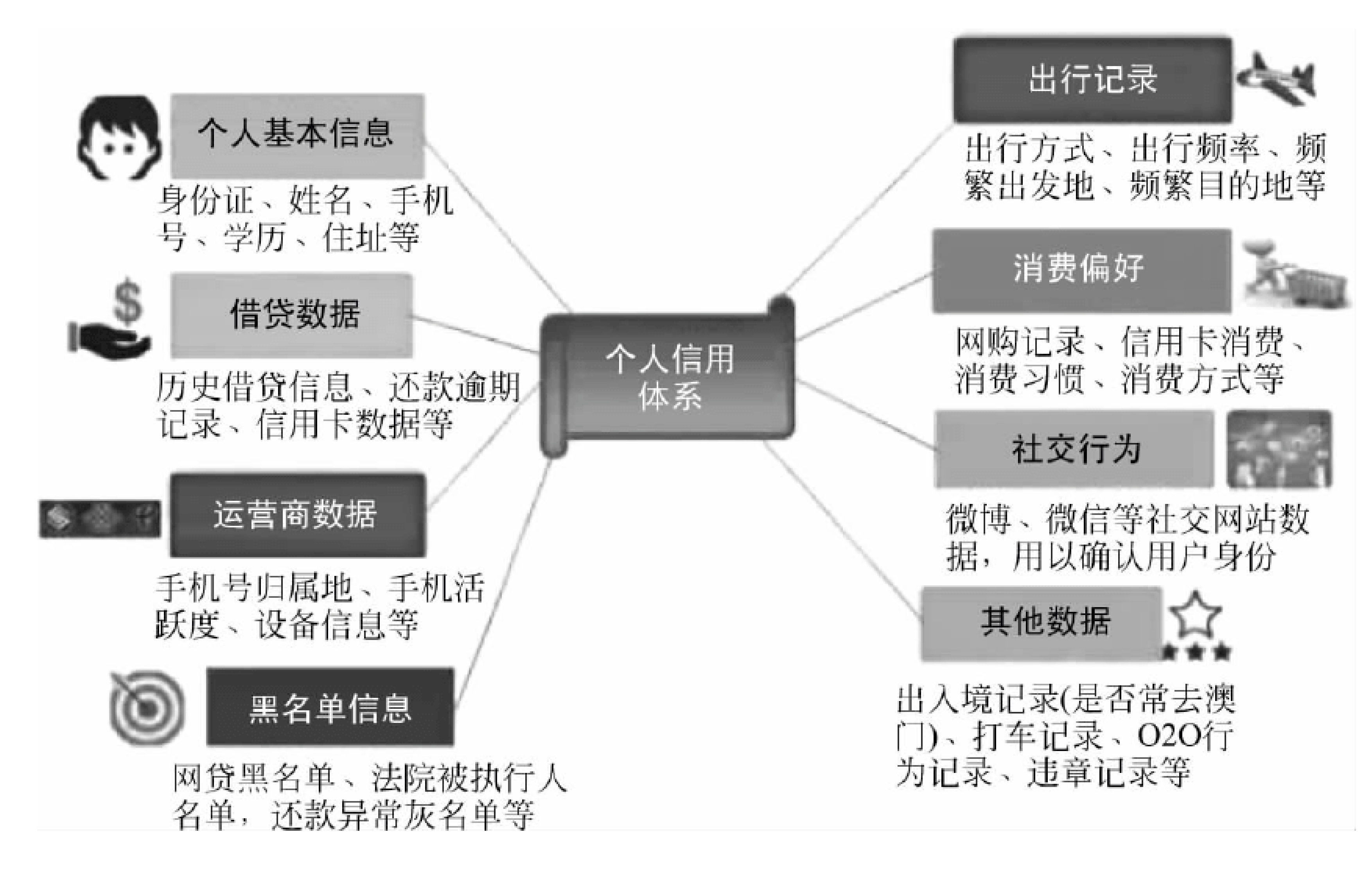


图 11-7　大数据个人征信体系

按照对个人信用风险影响的大小可以将个人信息分为强相关信息和弱相关信息，个人的姓名、身份证、手机号属于用户身份识别的强相关信息，借款用户的信用卡账单、月消费金额、网络购物真实流水分析等是用户还款能力的强相关信息，用户的历史借款记录、逾期笔数、借贷意图等是个人还款意愿的强相关信息。

用户其他的信息，例如用户的身高、体重、姓名、星座等信息，很难从概率上分析出其对用户个人信用的影响，这些弱相关信息，对用户的信用消费能力影响很小，可以忽略不计。

2. 大数据征信应用

(1) 大数据征信应用于个人信贷审批整个流程。个人信贷业务审批流程分为贷前审核、贷中决策、贷后管理三个部分，如图 11-8 所示。在贷前审核阶段，主要对借款人进行身份识别和信用评估，贷中决策阶段主要进行信用跟踪及风险预警，贷后管理阶段主要有逾期预警、失联修复、轨迹分析，信贷风险控制主要集中在贷前审核与贷后管理阶段。

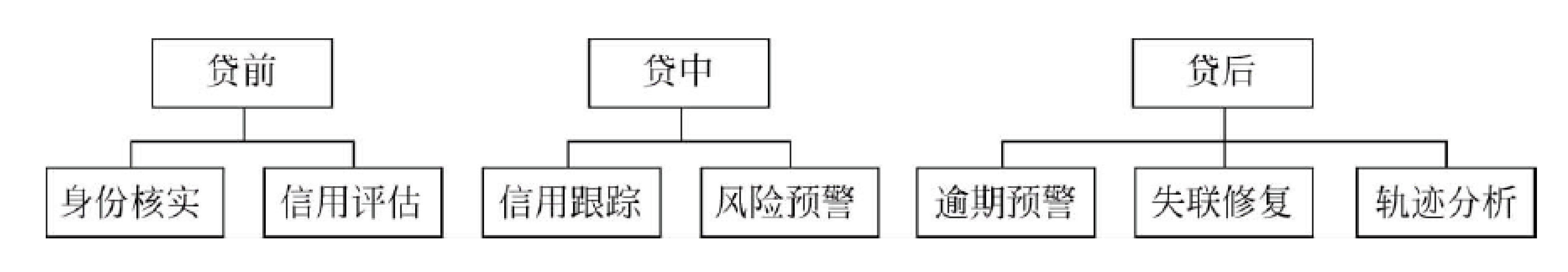


图 11-8 个人征信贷款审核图

贷前审核分为身份核实、信用评估两部分，由于个人信贷额度一般较小，因此对用户还款意愿的评估比还款能力的评估更为重要。

个人的姓名、电话、身份证件等人口属性信息主要用来对借款人进行身份识别，通过对借款人手机联系人的确认、居住地址位置、指纹、黑名单查询等来确定借款人身份是否真实，是否具有贷款资格，防止欺诈风险。

个人的历史借贷记录包括负债、是否逾期还款等信息，能够体现出个人负债情况，及信用度，负债额度高、恶意逾期还款次数较多的客户属于高风险客户；个人的消费数据包括借款用户的信用卡账单、月消费金额、网络购物真实流水分析可以对用户还款能力进行评估，具有高薪工作的用户且消费水平较高的客户，其贷款信用违约率较低；运营商数据可以对用户联系人、通话记录等进行分析，与贷款电话通话时间较长、换号频率高、经常关机的客户骗贷风险较高。

在贷中决策阶段，主要对用户进行信用跟踪及风险预警，实时监测信用主体的信用风险，例如，卷入法律纠纷、天灾人祸等，需及时做出风险预警。

贷后管理主要跟踪客户所属行业、客户经济状况、客户异常行为，包括其个人信用的变化，及时发现可能不利于贷款按时归还的问题，并提出解决问题的措施。举例来说，假如发现借款人在其他平台借款已经发生逾期、近期手机经常关机等迹象，则借款人有较高的概率逾期还款，需及时做出逾期预警；一旦客户已经失联，可以利用用户联系人、通话记录等进行分析，定位用户手机使用位置，了解到其联系人信息，结合出行记录等分析借款人行踪；利用借款人行踪、经济状况变化、消费等信息了解借款人逾期原因，是有钱不还还是因为经济能力等原因无钱可还，制定相应的催收方案。

(2) 大数据征信应用不仅限于传统金融机构，还可以与日常生活场景结合在一起。从

应用范围来看,目前大数据征信除了在金融机构、政府部门、公共服务等场景之外,还能与各类生活化、日常化的场景结合在一起,比如出行的租车免押金、住宿的入住免押金、购物的先试后买等各类日常履约场景相结合。

随着互联网,尤其是移动互联网的普遍化,人们的行为数据逐渐在互联网上沉淀,包括金融、餐饮、零售、旅游、社区、出行、教育、医疗、美容等诸多领域。新兴场景的出现,一方面,让征信走出常规的金融应用场景,扩大了个人征信的市场空间;另一方面,极大地提高了用户体验,进而提升了个人征信的使用黏性。

目前,不同机构数据资源共享仍然存在难度,因此不同的大数据征信产品侧重点不同,有的倾向电商信用行为,有的侧重互联网社交行为,有的反映借款人风险等。因此在全面评估个人信用风险时,可以结合多家机构的信用评估报告,从社交、电商、招聘、浏览行为、地理位置等不同角度对用户做出全息用户画像,判断其综合情况。

11.4.4 大数据反欺诈

1. 移动大数据的商业价值

在PC互联网时代,不管用户是否喜欢BAT,其网站仍然在那里。但是在移动互联网时代,如果一个用户不喜欢这个应用,就可以在两秒钟内删掉这个App,彻底中断和它的连接,无论其是不是BAT。在移动互联网时代,选择权完全转向用户,消费者将成为数字世界的中心。过去以品牌为中心的消费形式,将会转变为以消费者为中心的消费形式。

智能手机上安装的App和App使用的频率,可以代表用户的喜好。例如,喜欢理财的客户,其智能手机上一定会安装理财App,并经常使用;母婴人群也会安装和母婴相关的App,频繁使用;商旅人群使用商旅App的频率一定会高于其他移动用户。80后、90后将成为社会的主要消费人群,他们的消费行为将会以移动互联网为主,App的安装和活跃数据更加能够反映出年轻人的消费偏好。

智能手机设备的位置信息代表了消费者的位置轨迹,这个轨迹可以推测出消费者的消费偏好和习惯。在美国,移动设备位置信息的商业化较为成熟,GPS数据正在帮助很多企业进行数据变现,提高社会运营效率。在中国,移动大数据的商业应用刚刚开始,在房地产业、零售行业、金融行业、市场分析等领域取得了一些效果。

特别是在互联网金融领域的应用,移动大数据正在帮助互联网金融企业实施反欺诈,降低恶意诈骗给互联网金融企业带来的损失。

2. 恶意欺诈成为互联网金融的主要风险

近几年,互联网金融爆发式发展,2015年P2P的交易总额超过一万亿,将成为具有影响力的产业。近年来,大量的金融行业专业人士和传统产业资本进入到互联网金融领域,表明这个产业的生命力正在不断增强,有的P2P企业的年交易额已经突破百亿元,有的P2P企业估值也超过了15亿美金。

但是在P2P行业,其面对的风险也在加大,除了传统的信用风险,其外部欺诈风险正在成为一个主要风险。有的P2P公司统计过,带给P2P公司的最大外部风险不是借款人的坏账,而是犯罪集团的恶意欺诈。网络犯罪正在成为P2P公司面临的主要威胁之一,甚至在一些P2P公司,恶意欺诈产生的损失占整体坏账的60%。很多P2P公司将主要精力放在如何预防恶意欺诈方面。高风险客户识别和黑名单成为预防恶意欺诈的主要手段。

3. 移动大数据在反欺诈领域的应用

移动大数据中的位置信息代表了用户轨迹,商业应用较早。2014 年,美国移动设备位置信息的市场规模接近一千亿美金。但中国移动设备位置信息的商业应用才刚刚开始。

从技术上讲,定位移动设备的位置有三种方式,第一种是通过运营商的三个基站定位,其误差大概在 200m;第二种是通过手机 App 中的 GPS 位置信息定位,大概误差为 50m;第三种是通过 WiFi 定位,误差大概在 3～5m。在移动设备位置信息商业应用中,三种定位方式都被应用,室内以 WiFi 定位为主,室外以 GPS 定位为主。移动大数据在反欺诈领域具有以下应用场景。

(1) 用户居住地的辨别。线上的欺诈行为具有较高的隐蔽性,很难识别和侦测。P2P 贷款用户很大一部分来源于线上,因此恶意欺诈事件发生在线上的风险远远大于线下。中国的很多数据处于封闭状态,P2P 公司在客户真实信息验证方面面临较大的挑战。

移动大数据可以验证 P2P 客户的居住地点,例如,某个客户在利用手机申请贷款时,填写自己的居住地是上海。但是 P2P 企业依据其提供的手机设备信息,发现其过去三个月从来没有居住在上海,则这个人提交的信息可能是假信息,发生恶意欺诈的风险较高。

移动设备的位置信息可以辨识出设备持有人的居住地点,帮助 P2P 公司验证贷款申请人的居住地。

(2) 用户工作地点的验证。借款用户的工作单位是用户还款能力的强相关信息,具有高薪工作的用户,其贷款信用违约率较低。这些客户成为很多贷款平台积极争取的客户,也是恶意欺诈团伙主要假冒的客户。

某个用户在申请贷款时,如果声明自己是工作在上海陆家嘴金融企业的高薪人士,其贷款审批会很快并且额度也会较高。但是 P2P 公司利用移动大数据,发现这个用户在过去的三个月里面,从来没有出现在陆家嘴,大多数时间在城乡接合处活动,那么这个用户恶意欺诈的可能性就较大。

移动大数据可以帮助 P2P 公司在一定程度上来验证贷款用户的真实工作地点,降低犯罪分子利用高薪工作进行恶意欺诈的风险。

(3) 欺诈聚集地的识别。恶意欺诈往往具有团伙作案和集中作案的特点。犯罪团伙成员常常会集中在一个临时地点,雇佣一些人,短时间内进行疯狂作案。

大多数情况下,多个贷款用户在同一个小区居住的概率较低,同时贷款的概率更低。如果 P2P 平台发现短短几天内,在同一个 GPS 经纬度,出现了大量贷款请求,并且用户信息很相似,申请者居住在偏远郊区,这些贷款请求的恶意欺诈可能性就较大。P2P 公司可以将这些异常行为定义为高风险事件,利用其他的信息进一步识别和验证,降低恶意欺诈的风险。

移动设备的位置信息可以帮助 P2P 公司,识别出出现在同一个经纬度的群体性恶意欺诈事件,降低不良贷款发生概率。

4. 高风险贷款用户的识别

高风险客户也是 P2P 企业的一个风险。高风险客户定义比较广泛,除了信用风险,贷款人的身体健康情况也是一个重要参考。移动大数据的位置信息、安装的 App 类型、App

使用习惯，在一定程度上反映了贷款用户的高风险行为。

P2P 企业可以利用移动设备的位置信息，了解过去三个月用户的行为轨迹。如果某个用户经常在半夜两点出现在酒吧等危险区域，并且经常有飙车行为，这个客户定义成高风险客户的概率就较高。移动 App 的使用习惯和某些高风险 App 也可以帮助 P2P 企业识别出用户的高风险行为。如果用户经常在半夜两点频繁使用 App，经常使用一些具有较高风险的 App（例如某男同性恋应用），其成为高风险客户的概率就较大。

当用户具有以上的危险行为时，其身体健康就面临着较大的威胁，P2P 企业可以参考移动数据，提高将客户列为高风险客户的概率，拒绝贷款或者提前收回贷款，降低用户危险行为导致坏账的风险。

11.4.5　大数据精准营销

如今百货零售行业受到经济下行、线上电商的冲击、消费乏力而增长缓慢，行业竞争激烈。业态容易复制、商家品牌可以分享、推广活动没有新意等，真正学不来的是自身数据的处理、分析和挖掘，如何利用数据背后潜在的商业价值。

1. 大数据理解消费者行为特征

（1）供需精准化，大数据的第一个价值在于均衡供给和需求。

① 购物中心根据客流数量和历史数据告知各商家下个时段的预计顾客数，顾客 App 接收、蓝牙推送精准推荐的优惠券，引导顾客流量，均衡供需。

② 实现顾客标签管理的同时，把商家部分商品、套餐、服务数据化处理并且标签化，以便与目标顾客更精准匹配推荐。

③ “购物篮”式的精准化营销：将会员分为 15 个层级，为每一个层级推送完全不同但与之相应的信息。通过“云数据计算中心”为客户提供精准的个性化营销，管理层也能及时掌握每家商户的销售业绩以及市场变化状况及趋势。

④ 提供 WiFi 服务，将微信、微博、商家网站、App、往来、易信等连接成一个整体等，增加消费者的店内购物体验和购买转换率，让购物中心的全渠道零售管理逐渐从梦想成为可能。

（2）提升消费者体验，大数据让连接成本变低，能实时精准地把优惠推送给最有需求的人（例如，如果展厅某些场次观众很少，购物中心可向附近的会员发送免费参观券，用最小成本让顾客感受到意外惊喜和体验）。

① 根据大数据的消费客群、消费金额、消费频次、消费潜力分析，主动邀请高价值顾客和高影响力顾客成为 VIP 会员，为其提供预留车位、主动洗车、按摩椅贵宾室、一对一导购等特权服务。

② 顾客就是天生的、最好的推广员，口碑相传也是最好的营销广告……，引导顾客享受新服务，并引导他们随时在移动端提出感受和建议，并给以特别的惊喜和优惠，让其成为最好的推广宣传员。

③ 利用网络和数据进行一些有趣的游戏式活动促销。比如利用社交关系数据，提醒顾客他的朋友也在购物中心，双方碰面，对方加入 App，双方就都会有惊喜和奖励。又比如联合商家搞一些寻宝活动等。

④ 通过大数据可以分析出会员的行为习惯，消费额不同，购买商品差异，从而在某一时

间推送给会员某品牌的优惠券、O2O 活动或艺术沙龙等精准信息，从而实现大数据背后的精准化营销。

(3) 购物中心服务升级，个性化、精准化、人性化的服务是提高购物中心顾客黏性和依赖性的重要环节。

① 针对目前购物中心附近交通路线环境复杂、公共服务指向不清晰、商家变换快等情况，利用 App 和定位技术，提供导航路径服务。

② 针对购物中心周边文化展览单位汇集，提供路径导航，前期导览、导游等服务。

③ 提供购物中心周边车位的即时空位信息并给以路径导航。

④ 解决群体顾客的兴趣冲突，如一家三口到购物中心，孩子可以送到儿童乐园或培训教室，母亲可以去服装店购物，父亲可以去图书馆、运动馆……利用 App 的信息共享功能，家长可以随时关注、联络对方。

2. 以大数据构建线上线下高效运营平台

现在行业内众多的百货公司、购物中心、超市乃至专卖店都在使用客流监控系统，因此，购物中心也将综合利用先进的数据采集方法，采集更加全面准确的线下客流数据：蓝牙 4.0 信标、NFC 会员卡、WiFi 指纹技术、MEMS 顾客活动热力图、3D 传感＋视频监控＋人脸识别技术、LED 照明射频追踪技术、Euclid Zero 技术。

Euclid Zero 会识别出带 WiFi 配置的移动设备，并且不需要顾客自己接入商场的网络，就可以记录并分析客流情况。比如：有多少顾客、新老顾客占比、停留时间多长、到访频率如何、有多少是被橱窗内的海报或者摆设吸引而走进店里等数据。而这些数据可以帮助商家更全面地了解顾客群，进而优化服务策略、提升收益。

通过线下信息采集体系可以捕捉在广场里面所有的智能手机用户的行迹路线、所关注的商品和消费习惯，然后通过会员体系就可以掌握所有会员的各类信息和其特有的相关产品喜好。

建立购物中心实体店的线上 5D 全景购物中心，通过线上 5D 全景购物中心来挖掘线上客流并打通线上线下的交互经营，利用支付宝、易支付等解决支付问题。

线上可以通过 portal 页将用户导入该品牌的天猫店、支付服务窗、App、微信公众账号等。一旦导流系统完成，就可以通过 portal 页将实体店、天猫店、手机 App、企业支付账号和微博等互联网产品进行整合营销。与目前行业中广泛应用的简单 CPS 广告相比，前者的针对性更强，转化率更高。

当一位已注册的客人进入实体店，监控后台就能认出来，他过往的所有互动记录、喜好便会一一在后台呈现。通过对实体店顾客的电子小票、行走路线、停留区域的分析，来判别消费者的购物喜好，分析购物行为、购物频率和品类搭配习惯。

3. 利用大数据进行运营优化

(1) 优化会员生命周期管理：购物中心运营策略是立足于“经营客流”。单个消费者的单日消费轨迹追踪，利用价值并不高，而影响最大的是会员生命周期。通过对会员总体的生命周期管理，可以准确发现会员维护节点期、平台期、高价值消费期和预计的流失期——只有把握其中的规律，才有助于指导日常商业运营的会员管理。

(2) 精准获取消费者购物喜好。累积不同用户对品牌和折扣喜爱程度的数据，依托成

熟门店的相关数据，再根据新开门店所在城市的用户分析，可以导出新开门店组货和招商的指导意见。

① 商家销售经营数据库的建立。管理招商和科学精准的商铺定价；调整购物中心科学合理的业态配比。

② 商家销售经营数据库的管理。

③ 全维度数据分析体系：通过建立体系化分析矩阵，可以了解到经营业绩下降或增长的更深层原因，从而对症下药，对商户进行更加精准的扶持管理，从而实现更高的销售额，最终管理方获得更高的租金收益。

④ 商户经营扶持的业务平台：针对商户扶持管理的大数据业务平台，能够提供商户客流、销售、产品更新、展示、调价，甚至生产设计等各个方面的信息和预判指引，让商户从传统销售模式转为预测销售模式。

（3）会员消费行为数据库的建立。

① 通过对客户基础数据和消费数据的分析，将客户合理细分。

② 通过对客户多维度综合考量，充分挖掘客户价值，开展多种线下线上综合营销手段，达到精准营销、立体营销的目的，并节约营销推广成本。

③ 加强忠实会员的维系，借活跃会员口碑相传，提高品牌美誉度。

（4）会员消费行为数据库的管理。

全生命周期管理体系：与传统商业对会员管理只分析个体会员的单点指标，如个体会员的活跃度、消费情况等相比，消费者价值“全生命周期管理”理念，是基于对全体会员的研究。

通过对会员总体生命周期管理，可以准确发现会员的维护节点期、平台期、高价值消费期和预计的流失期。对即将进入维护节点期和流失期的会员，进行最大力度的维护管理，使其重新认识作为会员的价值所在。

（5）大数据技术的几个运用。

① 数据抓取：数据抓取作为大数据建设的基础，提供最广泛的数据来源。

② POS 系统管理每一家店铺的销售。

③ MIS 系统掌握每一天的销售变化。

④ 车流统计、客流统计和客流属性管理对应数据。

⑤ 客流管理：客流管理是对客流数据加以统计和分析，进行多维度研究。

⑥ App 管理跟踪服务。

⑦ CRM 社群：自建大数据体系，依托完善的经营数据和消费轨迹数据，精准分析并进行营销投放。整个 CRM 模式中，把消费者分成 15 个层级，每个层级都可以通过合理方法，进行精准推送，降低对顾客的骚扰程度，获取最大送达率。

⑧ 交互服务：建立 App 为消费者提供延伸服务，利用公众信息服务台和现场触摸自助设备，提供查询、导购、促销、优惠券及停车指导等服务；借助 iBeacon 技术，开展大流量的数据下载和产品推送服务。

11.5 金融大数据带来的产业变革

1. 机器学习快速发展，将会在金融风险管理领域广泛应用

数据科学家人才本身的供需关系将会朝着更加平衡的方向发展。在反欺诈和风控领域

将会使用更加成熟的技术来改善风控模型本身，并且加速发展实时分析监控和预警。这些快速的发展和变化会来自于业界领导者的传授和在现实世界的实践与应用。

2. 金融界大数据将引领产业发展和促进产业变革

每一年我们都能看到银行为了适应新技术而加大油门快速前进，同时在组织架构方面非常保守。业务和用户在 2016 年都将要激增而且会非常多变，结果就是在广阔的市场导致更强的可观察到的和可衡量的业务大量回归(不只是成本的下降)。

3. 数据治理合规性更加深入地集成到大数据平台

为了找到一个能够在合规性方面提供更强大功能的数据解决方案，许多银行都购买或者开发了单点解决方案，再不行就是用已经运行很多年的传统解决方案平台，但是这些解决方案都无法应对现今大规模爆发的数据。幸亏现在有越来越多的 Hadoop 改进方案来进行数据治理，改善血统和提供数据质量。更重要的是，这些新数据平台能够超越 Hadoop 平台达到传统数据存储的效果，并且做得更加大容量、更快，且在细节上达到合规性要求。此外在以后将继续看到为融合监管和风险控制(RDARR)中心服务的叫做“数据湖”方面的更多进展。

4. 金融服务业利用物联网向数据服务方面转变

这一波浪潮正是抓住大数据吸引力炒作/发力的好时机，同时金融服务应用的问题也很多。物联网数据在许多行业应用中已经实践(电信，零售，制造业)，这些行业驱动了物联网的数据的需求并且处于垄断地位。那么对于银行来说物联网数据是否能够用在 ATM 或者移动银行业务中？这些都是在多渠道实时数据流中值得探索的。例如，实时、多渠道的商业行为可以使用物联网数据对银行零售客户在正确的时间点提供适时的报价。或许我们反过来想想，金融公司可以将自己的服务内嵌到用户的某种“东西”或者设备或者其他和客户接触的点上，不在那些交易设施上，而是在家。

5. 软件供应商集市场、咨询和投资管理形成一个综合体

鼓吹与“从大数据获得更多利益”相关的新闻头条越奏越响。最终，这些观点都将被金融终端用户、可见的利益(或者不可见、无法衡量的利益)还有易用性等因素决定。大数据平台的建设核心将要提供的一个桥梁就是大数据，并且将其锐化突出。我们已经看到了市场数据供应商最喜欢的动作，但是并没有其他商业用户的应用，应朝这个方向努力(CRM，OMS/EMS 等)。

6. 风险控制和监管数据管理将继续成为顶级大数据平台的重要任务

增长与银行用户中心相关的商业行为将成为银行战略的重要举措，会有很多的银行把未来的战略与大数据关联起来。不论你的银行是不是基于发达的数据驱动的公司，朝着银行业务预测分析发展将是一条漫长的道路，会面临很大的挑战。同时也是一个必要的需求和被公司首席高官确认有意义的事。除非老天开恩或者监管机构放松要求，否则风险控制和监管仍然是下一年所有金融机构的首要挑战。

7. 金融服务业采用 Hadoop 作为关系型数据库进行存取将会大大增加

大家在不同的时间使用了相同的技术之间并没有任何差别。“长尾”效应还很遥远，但是中小型银行将会从 Hadoop 的以下几方面获益。

(1) 供应商将整合整套集成解决方案、服务、平台。

（2）用户社区持续成长，并能提供一个基础参考作为突破口。

（3）数据降载成为当今 Hadoop 一个“经典”应用（相对来讲），同时许多大数据专家继续在更大的数据集合上前进，未来将会有更多的普通人加入到大数据应用的行列。

8. 金融服务“大数据终结 App”理论在市场得到了越来越多的认可

FinTech 已经孵化了两三年，形成了大数据平台和用户间从前端到终端的连接。希望看到更多的银行作为证明概念来运行这些应用，这些实践将检验软件所提供的“完整解决方案”的基础。前端到终端和后端都应进行整合，而不是分割。大家可以看到市场迅速地从服务集扩展到后端，这将迎来银行业关于如何定位“大数据软件”和“传统软件”的激烈讨论。

9. 变化来了，获得前进动力的最后一次机会

随着越来越多的高可靠大数据平台的出现，安全专家、深层次的丰富元数据、集成 LEI 和其他标准成为一个严峻的现实。传统的数据方法是有效的，只是需要一些思想来充分利用新的解决方案——例如处理架构和数据建模。更深一层，随着大数据工作在前台，市场营销和风险控制方面形成的工作模式，我们能够看出这里面在办公的中后期业务上有明显和巨大的数据重叠部分，这些重叠能够很容易地应用在现有的数据湖中。我们预计，在中等的商业风险评估与性能相关的大数据的商业行为将迅速增加。更进一步，我们将看到关于如何切实带来后台功能的更深层次的交流（合作等）。

10. 银行的机构方将开始采用并从零售业务的方式来获取线索增进对于市场目标客户的了解

有一些纯 B2B 的公司利用大数据来改善客户商情，但是大部分时候他们处于 B2C 业务的不利地位，如信用卡业务、银行零售业、财富管理或者借贷业务。一个简单的跨界就是基金的配置（大型共同基金经理）从财富顾问网络和经纪人相互作用来改善数据收集的过程，同时也提高产品利用率。一旦被从客户群中移除，这对于共同基金通常是非常重要的，所以加强对于机构客户的理解显得尤为重要。

信任仍然是许多大型银行的使用新供应商“大数据”的主要因素。换句话说，当你展望下一年时，将会有很大的来自管理层的推动力，来把大数据项目移出 IT 然后放到商业用户手中。为了达成目的，我们需要考虑架构、功能、速度、可用性、安全性等问题。与往常一样，采用传统的严谨性以全新的架构布局并没有改变，传统架构将在成本和缓慢的进展中开始在新的 Hadoop 表现和融合的大数据的架构过程中逐步展现。

更进一步，将来一定会有更加强大的工具来处理现有的工作，例如数据治理、数据质量、参考数据管理、标准。这将要求各方持续的教育，即那些 IT 以外的继续教育，用以了解市场的快速发展。

最后，针对平衡开源和供应商解决方案将展开长期讨论。不是所有的开源项目设计之初就符合机构客户，开源项目传递了一种敏捷性需求开发——每个银行的需求都在不停变化，为大数据找到合适的点才是更加重要的。总而言之，2017 年的市场将会不断前行，混乱随之减少，同时会使大数据的海洋变得风平浪静。



中国制造大数据解决方案

我国的生产制造领域，已经从最初的粗放型粗加工阶段升级进入了精加工数控加工阶段，制造企业信息化程度已经有了很大提升，但是相对于发达国家高端制造业以及先进制造水平能力来说，我国制造企业还有很多工作要做，还有大量改进提高的空间。其中，大数据对制造领域的价值越来越被认可，相关的一些应用已经在逐步开展起来，成为助力我国制造业产业升级转型的一个重要推动力。大数据相关的技术应用与创新，已成为助力我国制造领域迈向高端制造高水平制造层面的迫切需求。

对于传统的制造业来说，随着企业信息化的逐步深入，数据积累到一定量之后，要想从这些数据中挖掘出更有价值的信息，来获得深刻的客户洞察，及时捕捉客户需求的变化趋势，就需要传统制造企业以客户为导向，了解客户的兴趣、偏好，通过各种渠道来获得用户对产品的反馈，需要处理好大数据，了解客户行为，将客户喜欢的产品及时交付。通过对大数据的获取、发掘和分析，企业可以更加经济地从多样化的数据源中获得更大价值，促进制造业按客户需求转型。生产设备实时监控数据，汇总分析辅助企业决策。在大型企业中，大量加工设备的实时运转情况汇总到一个平台，统计信息包括年度月度的设备忙闲时、加工磨损部件更换次数、设备出故障次数和原因等，这些信息汇总给企业决策层，以便了解生产加工密度，合理安排加工批次，合理接单，保持设备一直运转不至于空闲，总结关键部件磨损规律以合理安排备件更换，总结分析设备故障规律以合理安排预检维修。多方面的数据长期记录累加，便于企业利用数据分析有效因素，辅助决策大幅度提高生效率。

生产过程操作数据记录，汇总分析辅助工艺水平提升。在制造企业，对于加工过程的各类操作进行记录，并把记录关联到最终产品的质量上，通过长时间的阶段数据汇总，统计分析操作过程中的哪些环节会影响最终产品的质量，辅助加工企业进行工艺的调整改进，促进工艺水平的提升。

生产环境的监测数据，汇总分析辅助生产环境调整。对于某些加工过程中，环境因素影响最终产品质量和效果的情况，在生产中通过传感器实时收集温度、湿度、磁场强度等环境数据，汇总分析关联最终产品质量的因素，并不断调整相关因素再记录，最终找到能达到产品质量最高合格率的环境数据值，设定为最佳生产环境参数。

12.1　全球工业信息化发展历程和现状

西方发达国家是在基本完成工业化后，开始推进信息化的，其信息化是在成熟工业化的基础上发展起来的，因此在总体上呈现出先工业化、后信息化的梯度发展格局。信息与通信

技术已经并且还将继续成为促进发达国家经济增长强大的驱动器，因此，作为已经实现了工业化的国家，其信息技术应用和信息化所追逐的目标仍然包括传统产业的改造升级、新兴产业的发展，以及推动信息和知识的生产，否则就不可能保持他们在全球竞争中的“发达国家”地位。

纵观美欧日韩各国的发展战略，可以看出，各国在实施整体信息化进程中，均注重推进先进制造技术与信息科技技术的进一步融合，提供传统产业竞争力。总体上看，存在着两大推动力量：一是传统制造业借助信息化技术实施现代化的管理、设计和制造，从而提高生产、管理效率；另一个是将大量的信息化技术融入传统制造业产品流程，改进原有的产品制造过程，服务密集型导向的制造趋势日益明显。

12.1.1　美国工业信息化发展历程和现状

1. 美国工业信息化早期阶段

1946 年，美国福特公司的机械工程师哈德首先用“自动化”一词来描述生产过程的自动操作。1952 年，迪博尔德的《自动化》一书出版，他在书中认为“自动化”是分析组织和控制生产过程的手段。在 1952 年即商用电子计算机问世的第二年，美国柏森斯公司就以电子管元件为基础，设计了数控装置，试制了第一台三坐标数控铣床。

1974 年，也就是 Intel 公司第一个微处理芯片问世的第三年，第五代使用微处理芯片和半导体存储器的计算机数控装置研制成功。20 世纪 80 年代初，IBM 公司率先将计算机辅助设计(CAD)技术应用于产品设计。随后，计算技术的迅猛发展使得传统的自动化技术得到了全面的数字化改造，使产品研发、设计、生产、测试、供销等各个环节逐步实现智能化和网络化，信息化和工业化的融合进入了一个全新的发展时期。

2. 计算机集成制造

从 20 世纪 80 年代中期开始，美国大力提倡信息技术(当时主要是计算机技术，如网络、数据库、各种工业用的软件等)在制造业中的应用，目的是改变 20 世纪 70 年代因轻视制造业而造成的美国产品地位落后的状况，夺回生产优势。1973 年美国人约瑟夫·哈灵顿(J. Harrington)提出了 CIM(Computer Integrated Manufacturing)的概念，即计算机集成制造。哈灵顿认为，企业生产的组织和管理应该特别强调以下两个观点。

(1) 企业中的各种生产经营活动是不可分割的，是一个有机的整体，需要统一加以考虑。

(2) 整个生产制造过程实质上是信息的采集、传递和加工处理的过程。

CIM 是一种组织、管理与运行企业的理念，它将传统的制造技术与现代信息技术、管理技术、自动化技术、系统工程技术等有机结合，借助计算机技术、通信技术使企业产品全生命周期各阶段活动中有关人/组织、经营管理和技术三要素及其信息流、物流和价值流有机集成并优化运行，实现企业制造活动的信息化、智能化、集成优化，以达到产品上市快、高质、低耗、服务好、环境清洁，进而提高企业的柔性、健壮性、敏捷性，使企业赢得市场竞争。

3. 基于信息技术的敏捷制造

20 世纪 90 年代，美国根据本国制造业面临的挑战和机遇，为增强制造业的竞争能力和促进国家经济增长，克林顿总统提出了先进制造技术的 6 项行动。

其中的敏捷制造(Agile Manufacturing，AM)，是美国为恢复其在世界制造业的领导地位而提出的一种全新概念的生产方式，是美国在 21 世纪的制造战略。

敏捷制造是将柔性制造技术,熟练掌握生产技能的、有知识的劳动力以及促进企业内部和企业之间相互合作的灵活管理机制集成在一起,通过共同的基础设施,对迅速改变或者无法预见的消费者需求和市场机遇做出快速响应。

敏捷制造将制造系统空间扩展到全国,通过全美工厂网络建立信息交流的高速公路,建立全新的企业——“虚拟企业”或“虚拟公司”,以竞争能力和信誉为依据选择合作伙伴组成动态公司,进行企业大联合,共同冒险、共同获利。这是利用信息技术打破时空阻隔的一种新型企业,是一批为了完成某一特定任务,利用电子手段在短时间内迅速建立起灵活关系的合作者所构成的协作网络,不同于传统观念上的企业。如图 12-1 所示为敏捷制造跨企业合作模式。

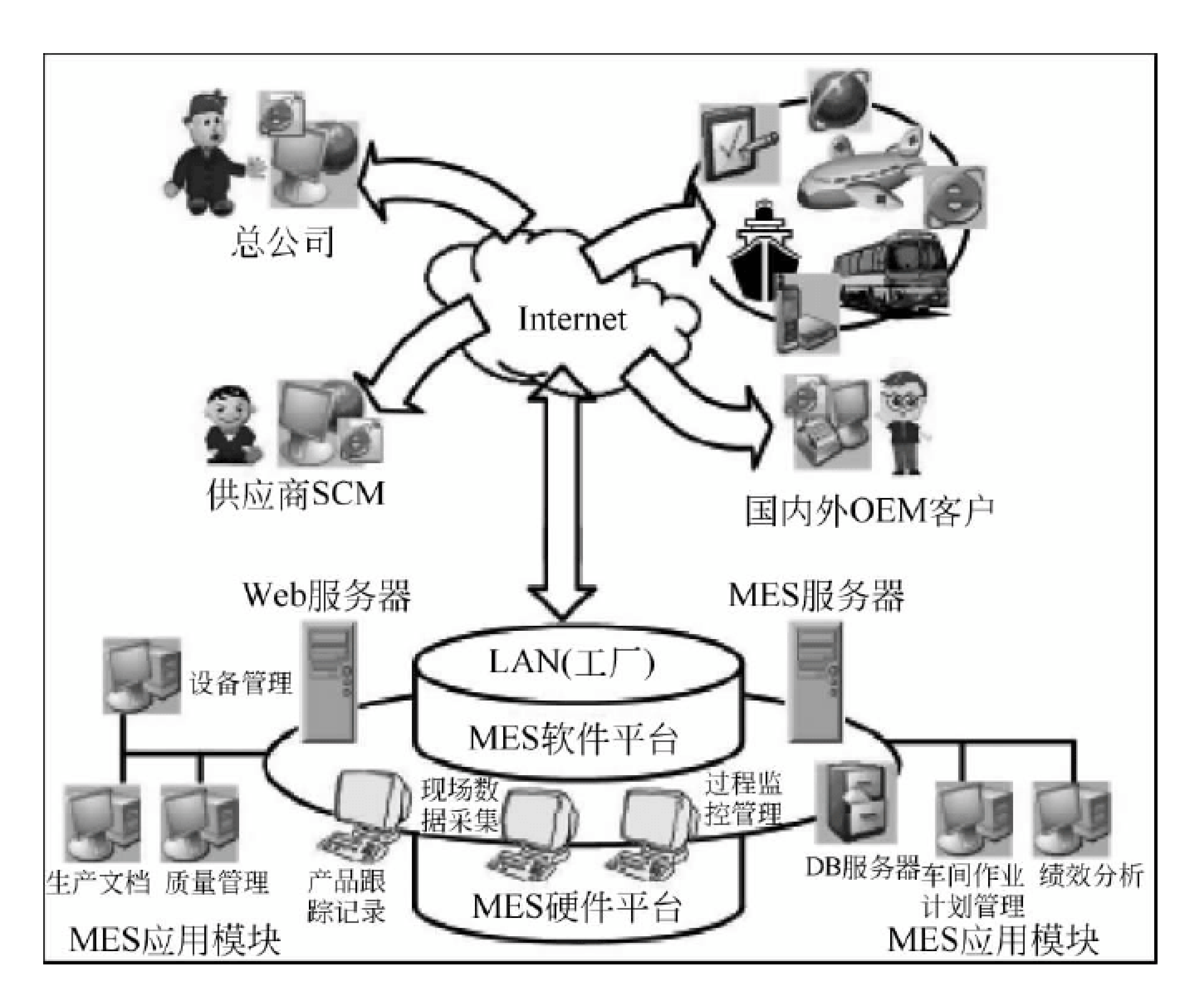


图 12-1 敏捷制造跨企业合作模式

4. 美国工业信息化现状

进入 21 世纪,美国相继发布《21 世纪信息技术计划》《网络与信息技术研究开发计划》和《网络空间安全国家战略》。美国是一个信息化领先于世界各国的国家,政府早在 1993 年就公布了《国家信息基础设施:行动纲领》。为了能够将美国的影响扩大到全球,美国在 1994 年 9 月提出建立“全球信息基础设施”的倡议,建议联通各国的国家信息基础设施,实现全国之间的信息共享。在 1997 年颁布的全球电子商务框架,鼓励在全球范围内促进电子商务发展。1998 年,美国麻省理工学院(MIT)的凯文·阿什顿(Kevin Ashton)在 Procter&Gamble 公司演讲中第一次提出物联网(Internet of Things)的概念,即通过在各种物体上增加射频身份识别或其他传感器,组成一个新的网络,并使现有的互联网步入一个新阶段。2002 年,美国出台网络空间国家安全战略,提出了 5 大优先发展领域和 47 项行动建议,将信息网络安全置于国家战略高度。

2008 年以来电子商务的快速发展也促进了产业不断细化,一些新型的电子商务模式呈现出良好的发展势头。2008 年 9 月,谷歌公司与通用电气公司对外宣布共同开发清洁能源业务,为美国打造国家智能电网。2009 年 1 月,IBM 首席执行官建议政府投资新一代的智

能基础设施,即“智慧地球”战略:将感应器嵌入和装备到电网、铁路、建筑、大坝、油气管道等各种物体中,形成物物相联,然后通过超级计算机和云计算将其整合,实现经济社会和物理世界融合。

总体来说,美国在推行信息化过程中注重在扩充全球势力范围和解决国内问题时的信息化导向,努力建设充满活力的“网络和信息技术生态环境”。另一方面,云计算方兴未艾,凭借其对全球数字信息资源的超强整合能力,进一步依托其高性能计算和云计算技术的领先优势,为全球客户提供崭新的海量数据密集型服务解决方案。

12.1.2 日本工业信息化发展历程和现状

1. 日本工业信息化早期阶段

20 世纪 60 年代以来,微电子半导体技术以及集成电路的发展,促进了日本电子与信息产业乃至整个工业信息化的发展。日本自 1963 年引进集成电路(IC)生产技术后,20 世纪 70 年代开发出大规模集成电路(LSI),20 世纪 80 年代进入超大规模集成电路(VLSI)时代。1970—1982 年,VLSI 以年均 50%的速度发展,使得整个电子产业的增长速度达到了 17%。同时,由于微电子半导体技术的迅速发展,集成电路的生产成本直线下降,带动整个制造产业的升级更新,引发了一场深刻的社会变革。

(1) 日本计算机产业的迅速崛起和壮大。伴随 LSI 技术水平的提高,计算机的性能越来越优异,价格也更低,日本计算机实现了超高速型和超小型化,并且不断从工业领域深入到家庭等社会应用。

(2) 日本产业机器人的广泛应用。由于微电子技术的飞跃进步,产业机器人的成本降低,推动了机器人向生产线的应用,不仅应用于工业领域,而且应用于农林、水产、矿业、医疗及第三产业等,极大地提高了生产效率。

(3) 汽车产业的发展。日本汽车制造业率先推广使用机器人自动生产线和计算机控制,使生产率大幅提高,质量提升且稳定,一举占据了世界市场的相当份额。

(4) 日本不断加快以微电子半导体技术为基础的计算机、数据图像传输处理、卫星通信、网络等信息技术产业的发展速度,并将信息技术及其产品应用到社会的各个领域,从生产到办公、家庭,迅速提升了整合社会的信息化程度。

1995 年,日本东京大学成立了机械制造信息学系,开始重视制造信息学在制造系统中的地位和作用,并开展了相关研究。正是由于日本政府重视信息技术的投资,才使日本的信息产业得以快速发展。

2. 智能制造系统项目

日本在 1991 年 1 月发起了智能制造系统(Intelligent Manufacturing System,IMS)的国际合作研究开发计划。该项计划旨在组合工业发达国家的先进制造技术,包括日本的工厂与车间的专业技术、欧洲共同体的精密工程专业技术和美国的系统专业技术;探索将研究成果转化为生产技术的途径;开发下一代的标准化技术。其重点是实现当前生产技术的标准化,开发出能不受生产环境和国界限制、彼此合作的高技术生产系统。通过各发达国家的共同研究,制造业在接受订货、产品开发和设计、生产、物流直至经营管理的全过程中,做到装备生产线的自律化,并实现自律化的装备和生产线在系统整体上的协调和集成,由此来适应制造活动全球化的发展趋势,减少过于庞大的重复投资,并通过先进、灵活的制造过程

来解决制造系统中的人因问题。“自律化”是指能够根据周围环境以及生产作业状况自主地进行判断并采取适当的行动。欧美有许多国家参加了这一计划。人们对智能制造系统信息技术的作用给予了很高的期望。虽然最终还是没有实现完全的智能制造，但对推进制造系统超智能化方向发展起到了重要的作用。

12.1.3 德国工业信息化发展历程和现状

1. 德国工业信息化概述

欧盟制造业因为长期以来形成的产业文化，供应商、制造企业、服务企业和用户之间业已建立了相互联系的广泛网络。其成员国拥有一流的研究能力，可以产生高水平的知识，具有良好的科学素养。另外，欧盟制造业99%的企业都是中小企业，具有很好的适应性、灵活性、创新能力和企业家精神，更有利于促进和实现企业之间合作竞争。欧洲较早实现了可持续发展战略，对环境保护、清洁生产以及环境友好生产过程的大量投资，已经形成了新的制造模式。经历了20世纪60年代自动化制造阶段和20世纪70年代精益制造后日本企业崛起带来的激烈冲击，欧盟各国一直在考虑如何更好地利用信息技术来建设信息经济社会，增强包括制造业在内的竞争力。为了提升欧洲制造业竞争力，欧盟委员会邀请来自研究机构和产业界的专家，经过讨论形成了指导未来欧洲制造业发展的《未来制造业：2020年展望》，并于2004年12月在荷兰恩斯赫德市(Enschede)召开的未来的制造业(Manufuture)会议上发布。在欧盟内部，与企业信息化建设有关的政府性机构和组织主要有欧盟委员会下的企业与工业总司、信息社会和媒体总司、欧洲信息中心和欧洲经济和社会委员会。性质不同，职责也不同，但他们在推进企业信息化建设方面相互协调、相互补充，各个机构的职责、出台的政策规划和举办的重大活动很少出现重复的情况，而是相互补充和相互支持。

2. 德国的“生产2000”计划

“生产2000”(Producktion 2000)计划是由德国政府、企业界、科技界和工会组织共同提出的一项战略计划。该项目总共投资4.5亿马克，执行时间为1995—1999年。

“生产2000”计划的研究重点如下。

(1) 产品的开发方法和制造方法，特别要研究如何缩短产品开发和产品制造的周期，以便对新的市场需要做出快速响应；

(2) 产品制造过程中的经济学，即开发可重复利用的材料并制定新材料的标准，开发可重复利用的产品，开发能进行“清洁制造”的制造过程；

(3) 面向制造的后勤学，特别是研究加速产品制造过程和减少运输费用的方法，同时也应考虑减少对环境的负面影响；

(4) 面向制造的信息技术，特别要研究通信技术，开发面向制造的高效的、可控的系统；

(5) 在“动荡”环境中的生产，即研究开放的、具有学习能力的生产组织结构，提高对市场变化的响应速度；

(6) 其他热门课题，如全球制造、企业协作和与其有关的标准。

3. i2010战略计划

2005年6月，欧盟委员会在比利时布鲁塞尔公布了一个新的战略计划——《i2010战略计划：欧洲信息社会2010》(i2010-Initiative：European Information Society)，其目的在于促

进欧盟经济增长和创造就业。i2010 战略计划是继 2000 年欧洲理事会制定的里斯本战略目标"到 2010 年把欧洲建设成世界上经济最活跃,最有竞争力的知识经济体"后,提出的又一个重要的战略计划,是欧盟为了应对现代信息社会的巨大挑战的一个产物。其为欧盟信息化的发展设定了三个目标:①建设一个统一的欧洲信息空间,向用户提供在价格上可以承受的、安全的高宽带通信以及内容丰富的、多样化的、数字化的服务;②在现代信息技术的研究和创新中,要有世界水平的表现,以缩小欧洲与其竞争对手之间的差距;③建设一个包容性的信息社会。

4. 欧盟物联网行动计划

2009 年 6 月,欧盟委员会向欧盟提交了《欧盟物联网行动计划》,以确保欧洲在构建物联网的过程中起主导作用。该计划描绘了物联网技术应用的前景,并提出要加强欧盟对物联网的管理,消除物联网发展的障碍。该行动计划提出以下建议:加强物联网管理;完善隐私和个人数据保护;提高物联网的可信度、接受度和安全性;推广物联网标准化;加强相关系统与关键技术研发;建立开放式的创新环境;增强机构间协调;加强国际对话;推广物联网标签、传感器在废物循环利用方面的应用;加强对物联网发展中的无线频谱与电磁影响的监测、统计和管理。而 2004 年发布的《未来制造业:2020 年展望》报告,基于研发和创新的发展战略,强调从个人竞争转向系统竞争,标准的 ICT 接口;参与虚拟工程和虚拟制造伙伴的开放网络,积极采用新的商业模式。从欧盟内部来看,不同成员国之间企业的信息化水平存在巨大差距,数字鸿沟明显。德、法、英等西欧国家企业信息化水平较高,而一些东欧国家的信息化水平却比较落后。

5. 德国工业 4.0

2013 年 4 月的汉诺威工业博览会上"工业 4.0"项目被正式推出。为了在新一轮工业革命中占领先机,在德国工程院、弗劳恩霍夫协会、西门子公司等德国学术界和产业界的建议和推动下,这一研究项目是 2010 年 7 月德国政府《高技术战略 2020》确定的十大未来项目之一——旨在支持工业领域新一代革命性技术的研发与创新。在工业科研联盟的倡议下,在工业 4.0 平台上的合作伙伴们已经为自己确立目标,贯彻德国政府的战略举措,以确保德国工业的竞争力。从本质上讲,工业 4.0 包括将虚拟网络—实体物理系统技术一体化应用于制造业和物流行业,以及在工业生产过程中使用物联网和服务技术。这将对价值创造、商业模式、下游服务和工作组织产生影响。

工业 4.0 计划具有巨大潜力主要表现在以下几个方面。

(1) 满足用户个性化需求。工业 4.0 允许在设计、配置、订购、规划、制造和运作等环节能够考虑到个体和客户的特殊需求,而且即使在最后阶段仍能变动。在工业 4.0 中,有可能在一次性生产且产量很低(1 批量)的情况下仍能获利。

(2) 灵活性。基于 CPS 的自组织网络可以根据业务过程的不同方面,如质量、时间、风险、鲁棒性、价格和生态友好性等,进行动态配置。这有利于原料和供应链的连续"微调"。也意味着工程流程可以更加灵活,制造工艺可以被改变,暂时短缺(例如供应问题)可以得到补偿,输出的大量增加可以在短时间内实现。

"工业 4.0 为德国提供了一个机会,使其进一步巩固其作为生产制造基地、生产设备供应商和 IT 业务解决方案供应商的地位。令人鼓舞的是,我们可以看到德国的所有利益

相关方在紧密合作，通过工业 4.0 平台，一起向前迈进，加以实施。”——孔翰宁（Henning Kagermann）

(3) 决策优化。为了在全球市场上取得成功，在短时间内能够做出正确决定变得越来越关键。工业 4.0 提供了端到端的实时透明，使得工程领域的设计决策可以进行早期验证，并且既可以对干扰做出更灵活的反应，还可以对生产领域中公司的所有位置进行全局优化。

(4) 资源生产率和利用效率。工业制造过程的总体战略目标仍然适用于工业 4.0：在给定资源量（资源生产率）的前提下，得到尽可能高的产品输出；使用尽可能低的资源量，达到指定的输出（资源利用效率）。CPS 在贯穿整个价值网络的各个环节基础上，对制造过程进行优化。此外，系统可就生产过程中的资源和能源消耗或降低排放进行持续优化，而不是停止生产。

(5) 通过新的服务创造价值机会。工业 4.0 开辟了创造价值的新途径和就业的新形式，比如通过下游服务。智能算法可用于各种大量数据（大数据），这些数据是为了提供创新服务而由智能设备所记录的。尤其是对于中小企业和初创公司来说，有显著的机遇发展 B2B（企业对企业）服务。

(6) 应对工作场所人口的变化。通过工作组织和能力发展计划相结合，人与技术系统之间的互动合作将为企业提供新的机会，将人口变化转化为自身的优势。面对熟练劳动力的短缺和日益多样化的劳动力（如年龄、性别和文化背景），工业 4.0 将提供灵活多样的职业路径，让人们的工作生涯更长，并且保持生产能力。

(7) 工作和生活的平衡。使用 CPS 的公司更加灵活的工作组织模式，意味着它们可以很好地满足员工不断增长的需求，让员工在工作与私人生活之间，以及个人发展与持续的职业发展之间实现更好的平衡。例如，智能辅助系统将提供新的组织工作的机会，即提供一种灵活的新标准以满足公司的需要和员工个人的需求。随着劳动力规模的缩减，CPS 公司在招聘最优秀员工方面将具备明显优势。

(8) 高工资仍然具有竞争力。工业 4.0 的双重战略将使得德国保持供应商的领先地位，并且成为工业 4.0 解决方案的主导市场。

然而，工业 4.0 不会对相关行业构成纯技术层面或与信息技术相关的挑战。不断变化的技术也将会对组织方面带来深远影响，它提供了开展创新的商业和企业模式、提高员工参与度的机会。20 世纪 80 年代初，通过将可编程逻辑控制器（PLCs）应用于制造技术，使制造自动化更加灵活。与此同时，通过采用一种基于社会伙伴关系的方法、管理对劳动力的影响，德国成功地进行了第三次工业革命。德国强大的工业基础、成功的软件产业和在语义技术方面的诀窍意味着德国可以很好地实施工业 4.0。德国有可能克服目前的障碍，如技术接受问题或劳动力市场熟练工人数量有限的问题。然而，只有所有利益相关方共同努力，挖掘物联网和服务为制造业带来的潜力，才有可能确保德国工业的未来。自 2006 年以来，德国政府已在其高技术战略下推动物联网和服务。一些技术项目已经成功启动。工业科学研究联盟正在利用工业 4.0 计划跨部门推进这一举措。在执行过程中下一步顺理成章的是建立第四次工业革命平台，由德国信息技术、通信、新媒体协会（BITKOM）、德国机械设备制造业联合会（VDMA）以及德国电气和电子工业联合会（ZVEI）三个专业协会共同建立秘书处。下一步的任务就是为关键的优先主题制定研发路线图。确保德国制造业的未来——这是工业 4.0 平台的合作伙伴确立的目标。该平台邀请所有相关的利益方继续探索工业 4.0

带来的机遇，只有这样，才可以帮助确保成功实施工业 4.0 的革命前景。

12.1.4 我国工业信息化发展历程和现状

三十多年来，我国基本形成了一条符合国情的工业信息化道路。与国外水平相比，在应用方面，特别是高端应用方面差距不大，但是应用面还比较窄；支撑工业应用的信息化产品尚处于初级发展阶段，差距较大，有的甚至非常大。

我国工业信息化发展历程大致可以归结为以下 6 个方面。

1. 国家 863 计划 CIMS 主题

1986 年国家 863 计划（高技术研究发展计划）开始论证和实施，首先启动对我国经济发展有重大影响的 7 个高技术领域公关。其中的自动化技术领域包括计算机集成制造系统（CIMS）主题和智能机器人主题。

1994 年和 1999 年，国家 CMIS 工程中心和华中理工大学分别荣获美国制造工程师学会的“大学领先奖”；1995 年，北京第一机床厂的 CMIS 工程又获美国制造工程师学会的“工业领先奖”。

2. “九五”“甩图板”工程

“九五”初期，当时的国家科委（中华人民共和国科学技术委员会）主任宋健同志提出“甩图板”的口号。虽然“甩图板”只是一个历史意义的突破口，远不是制造业信息化的最终目标和最高境界，但却形象地描述了“九五”CAD 推广应用工程的阶段性愿景在全国产生巨大的号召力。20 世纪 90 年代的“甩图版”“扔算盘”工程，推动了企业信息化普及高潮：在六百多家企业进行了 CAD 技术示范应用，三千多家企业进行了重点应用，并带动了万家企业开发 CAD 应用。

3. “十五”国家制造业信息化工程

图 12-2 为“十五”期间制造业信息化工程建设任务体系结构。其特点是形成完善的工程体系，省市协同推进。国家还成立了制造业信息化工程协调领导小组。

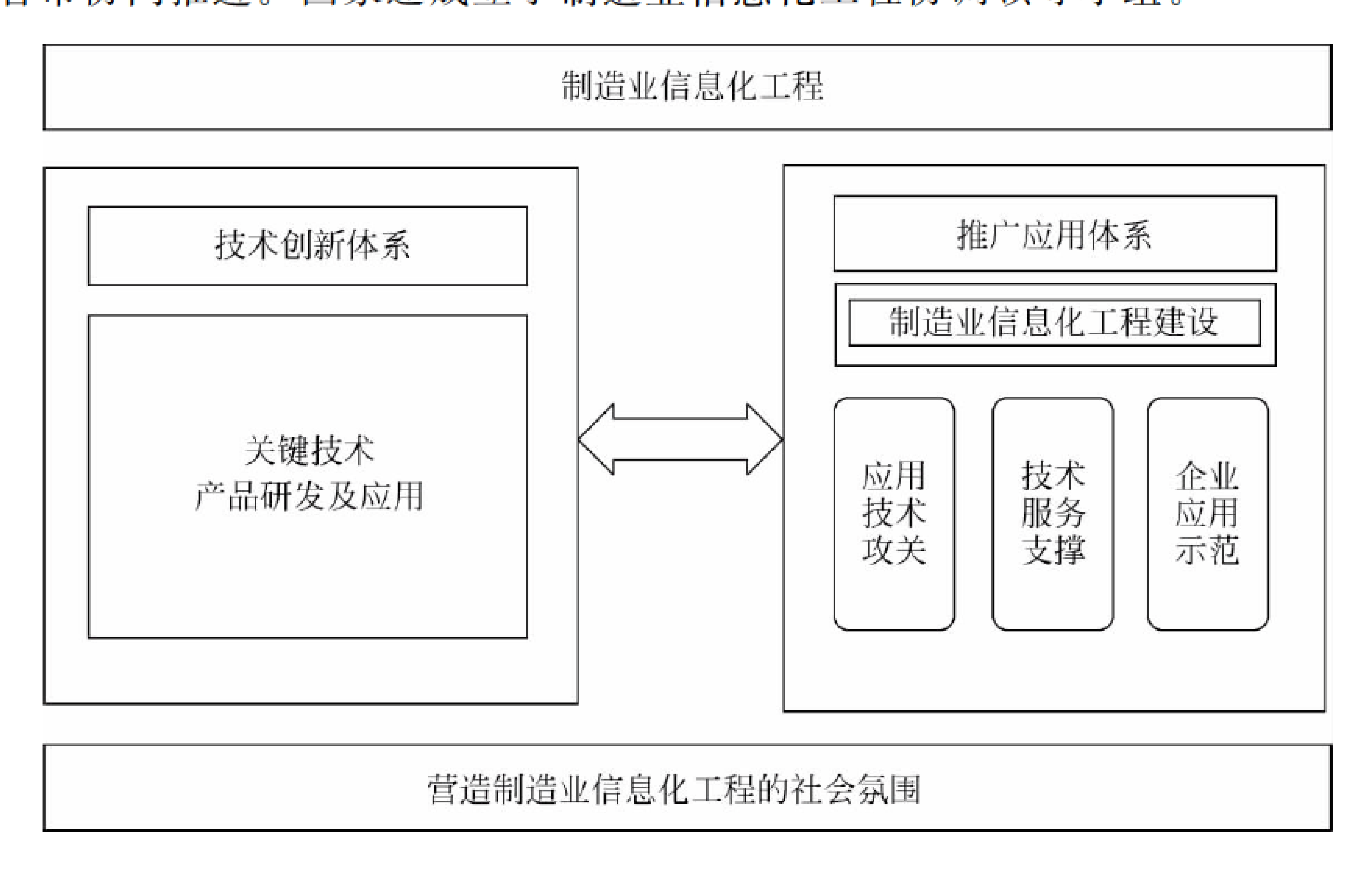


图 12-2 “十五”期间制造业信息化工程

“十五”期间培育了一批制造业信息化专业服务机构，在27个省、49个重点城市的六千多家企业推广了制造业信息化工程。

4. “十一五”国家“甩图纸”“甩账表”工程

“十一五”期间，科技部提出组织制造业企业实施设计制造一体化的“甩图纸”示范推广工程和经营管理信息化的“甩账表”示范推广工程，简称“两甩”工程，这成为“十一五”制造业信息化科技示范工程预期实施效果的阶段性愿景。

“十一五”期间，以集团企业、骨干企业、中小企业为对象，以集成与协同为重点，向纵深方向推进制造业信息化发展，全面提升了制造企业核心竞争力。

5. “十二五”国家制造业信息化工程

“十二五”期间国家科技部门将以服务为手段，以增效为目标，以升级转型为标志，继续大力推动我国制造业信息化工程。

6. “十三五”国家制造业信息化工程

“十三五”期间，提出《中国制造2025》，主要内容包括：坚持“创新驱动、质量为先、绿色发展、结构优化、人才为本”的基本方针，坚持“市场主导、政府引导，立足当前、着眼长远，整体推进、重点突破，自主发展、开放合作”的基本原则，通过“三步走”实现制造强国的战略目标：第一步，到2025年迈入制造强国行列；第二步，到2035年我国制造业整体达到世界制造强国阵营中等水平；第三步，到新中国成立一百年时，我国制造业大国地位更加巩固，综合实力进入世界制造强国前列。

12.1.5 我国《中国制造2025》的发展战略

制造业是国民经济的主体，是科技创新的主战场，是立国之本、兴国之器、强国之基。当前，全球制造业发展格局和我国经济发展环境发生重大变化，必须紧紧抓住当前难得的战略机遇，突出创新驱动，优化政策环境，发挥制度优势，实现中国制造向中国创造转变，中国速度向中国质量转变，中国产品向中国品牌转变。

深入实施《中国制造2025》，通过政府引导、整合资源，实施国家制造业创新中心建设、智能制造、工业强基、绿色制造、高端装备创新等5项重大工程，实现长期制约制造业发展的关键共性技术突破，提升我国制造业的整体竞争力。

围绕实现制造强国的战略目标，《中国制造2025》明确了9项战略任务和重点：一是提高国家制造业创新能力；二是推进信息化与工业化深度融合；三是强化工业基础能力；四是加强质量品牌建设；五是全面推行绿色制造；六是大力推动重点领域突破发展，聚焦新一代信息技术产业、高档数控机床和机器人、航空航天装备、海洋工程装备及高技术船舶、先进轨道交通装备、节能与新能源汽车、电力装备、农机装备、新材料、生物医药及高性能医疗器械等十大重点领域；七是深入推进制造业结构调整；八是积极发展服务型制造和生产性服务业；九是提高制造业国际化发展水平。

1. 全面提升工业基础能力

实施工业强基工程，重点突破关键基础材料、核心基础零部件(元器件)、先进基础工艺、产业技术基础等“四基”瓶颈。引导整机企业与“四基”企业、高校、科研院所产需对接。支持全产业链协同创新和联合攻关，系统解决“四基”工程化和产业化关键问题。强化基础领域

标准、计量、认证认可、检验检测体系建设。实施制造业创新中心建设工程，支持工业设计中心建设。设立国家工业设计研究院。

2. 加快发展新型制造业

实施高端装备创新发展工程，明显提升自主设计水平和系统集成能力。实施智能制造工程，加快发展智能制造关键技术装备，强化智能制造标准、工业电子设备、核心支撑软件等基础。加强工业互联网设施建设、技术验证和示范推广，推动“中国制造＋互联网”取得实质性突破。培育推广新型智能制造模式，推动生产方式向柔性、智能、精细化转变。鼓励建立智能制造产业联盟。实施绿色制造工程，推进产品全生命周期绿色管理，构建绿色制造体系。推动制造业由生产型向生产服务型转变，引导制造企业延伸服务链条、促进服务增值。推进制造业集聚区改造提升，建设一批新型工业化产业示范基地，培育若干先进制造业中心。

3. 推动传统产业改造升级

实施制造业重大技术改造升级工程，完善政策体系，支持企业瞄准国际同行业标杆全面提高产品技术、工艺装备、能效环保等水平，实现重点领域向中高端的群体性突破。开展改善消费品供给专项行动。鼓励企业并购，形成以大企业集团为核心，集中度高、分工细化、协作高效的产业组织形态。支持专业化中小企业发展。

4. 加强质量品牌建设

实施质量强国战略，全面强化企业质量管理，开展质量品牌提升行动，解决一批影响产品质量提升的关键共性技术问题，加强商标品牌法律保护，打造一批有竞争力的知名品牌。建立企业产品和服务标准自我声明公开和监督制度，支持企业提高质量在线检测控制和产品全生命周期质量追溯能力。完善质量监管体系，加强国家级检测与评定中心、检验检测认证公共服务平台建设。建立商品质量惩罚性赔偿制度。

5. 积极稳妥化解产能过剩

综合运用市场机制、经济手段、法治办法和必要的行政手段，加大政策引导力度，实现市场出清。建立以工艺、技术、能耗、环保、质量、安全等为约束条件的推进机制，强化行业规范和准入管理，坚决淘汰落后产能。设立工业企业结构调整专项奖补资金，通过兼并重组、债务重组、破产清算、盘活资产，加快钢铁、煤炭等行业过剩产能退出，分类有序、积极稳妥处置退出企业，妥善做好人员安置等工作。

6. 降低实体经济企业成本

开展降低实体经济企业成本行动。进一步简政放权，精简规范行政审批前置中介服务，清理规范中介服务收费，降低制度性交易成本。合理确定最低工资标准，精简归并“五险一金”，适当降低缴费比例，降低企业人工成本。降低增值税税负和流转税比重，清理规范涉企基金，清理不合理涉企收费，降低企业税费负担。保持合理流动性和利率水平，创新符合企业需要的直接融资产品，设立国家融资担保基金，降低企业财务成本。完善国际国内能源价格联动和煤电价格联动机制，降低企业能源成本。提高物流组织管理水平，规范公路收费行为，降低企业物流成本。鼓励和引导企业创新管理、改进工艺、节能节材。

12.2 工业信息化技术集成和协同发展方向

集成和协同构成了工业信息化技术发展主旋律，它既是实现各种工程系统必不可少的技术，也是带动各种单元技术发展的动力。集成和协同不是简单地将两个或多个单元联系在一起，而是将原来没有联系或者联系不紧密的单元有机地组成为有一定功能的、相互间紧密联系的新系统，从而产生更大的效益。

在工业信息化技术发展不同阶段，集成和协同有不同的内涵和外延，图12-3分别从集成和协同的空间跨度、集成和协同的时间跨度、集成和协同的重点、集成和协同的对象以及主要集成和协同技术等方面表示了集成和协同技术的发展过程。

集成和协同的空间跨度	部门内	企业内 部门间	企业间 供应链管理 流程工业综合自动化 网络化制造等
集成和协同的时间跨度	产品制造过程中不同阶段	产品制造过程	产品全生命周期管理PLM，制造服务等
集成和协同的重点	信息	过程	知识 知识管理 智能制造
集成和协同的对象	几何模型	几何模型+ 部分性能模型	多学科模型 多学科设计优化 MDO等
主要集成和协同技术	LAN CAD/CAM/CAPP	Internet/Intrannet ERP/PDM 数据库，DCS	企业集成 物联网 云计算等

图12-3 工业信息化集成和协同发展过程

12.2.1 集成和协同的空间跨度

从集成和协同的空间跨度来看，已经从原先的部门内、企业内各部门间，发展到追求整个增值链效益最大化的企业间集成和协同，目前的代表技术有供应链管理、流程工业综合自动化和网络化制造等。

供应链管理(Supply Chain)是现代物流中供应、分配和销售渠道及过程一体化管理的结果，涵盖所有参加供应、生产、分配和销售过程的企业，是现代物流活动中的核心过程和主线。以跨组织、连续性等为特征的供应链集成是现代物流管理的核心理念，是系统化

和系统整体性的体现以及现代社会发展的客观要求。集成供应链管理技术遵循融合(Syncretism)、共生(Symbiosis)和协同(Synergy)的“3S”原则,保持供应链系统的高效性和灵活性,从而保证整个供应链的成长性和持续发展。正在兴起的敏捷供应链(Agile Supply Chain)充分体现了这种集成化的思想,代表了工业企业物流系统管理的最新发展方向之一。

流程工业综合自动化是将先进的工艺技术、现代管理技术和以先进控制及优化技术为代表的信息技术相结合,将流程工业企业的经营管理、生产过程控制、运行作为一个整体来进行综合的管理,将 ERP(Enterprise Resource Planning)/MES(Manufacturing Execution System)/PCS(Process Control System)三级结构应用于流程工业企业。其中,PCS 是信息处理和控制的基础;MES 以生产调度为核心,起着承上启下的关键作用;ERP 则是以资源的优化配置、调度和经营决策为目标的管理层。应用多智能体等信息技术,从生产过程的全局出发,将生产加工技术与现代管理技术有机集成,形成一个集控制、监测、优化、调度、管理、经营和决策等功能于一体的协同递阶控制系统,实现企业/企业间的优化运行、优化控制和优化管理,从而形成适应各种生产环境和市场需求、总体最优、高质量、高效益、高柔性的现代化企业综合自动化系统。

12.2.2 集成和协同的时间跨度

从集成和协同的时间跨度来看,已经从原先仅考虑产品生命周期的某一阶段,发展为产品全生命周期管理,目前的代表技术是产品生命周期管理和制造服务技术。

经济全球化和信息技术的快速发展,使工业企业的竞争环境、发展模式及活动范围等发生了深刻的变化。在这种背景下,产品生命周期管理(Product Lifecycle Management,PLM)应运而生。从发展的趋势来看,PLM 正在迅速地从一种竞争优势转变为参与竞争所必须具备的技术。产品生命周期管理是一种在系统思想指导下,利用计算机技术、管理技术、自动化技术和现代制造技术等对产品全生命周期管理内与产品相关的数据、过程、资源和环境等进行管理。通过实施 PLM,企业各部门的员工、最终用户和合作伙伴等可以高效地协同工作,最终产品能达到综合最优。产品生命周期管理系统是一种面向数据、资源和过程的产品技术信息化集成系统。PLM 解决方案涵盖了从市场需求分析、开发设计、测试验收、生产制造、安装、运行、维护、服务以及报废回收等产品的整个生命周期(图 12-4)。从技术角度来看,PLM 的逐渐广泛应用与 PDM 技术的成熟和深化具有十分密切的联系,目前这两种技术还在不断发展之中,并将得到越来越广泛的应用。

经济全球化、信息技术的革命和现代管理思想的发展,使得全球制造业发生了重大变化。

同质化的竞争和供大于求的市场,使企业原有的生产、技术和资金等优势越来越不明显,产品利润率日益降低。发达国家跨国制造企业纷纷实施归核化战略和差异化战略,进行产品创新和服务创新,将经营重点放到核心业务价值链中本身优势最大的环节上,通过实施战略性外包增强差异性竞争优势。这就使原本完整连续的制造业产业价值链断裂分解,与渗透进来的服务业价值链混合,实现了制造业与现代服务业的产业融合,产生了全新的现代制造服务业价值链。因此,制造业已不仅仅提供产品,而是提供产品、服务、支持、自我服务和知识的“集合体”,制造业企业正在转变为某种意义上的服务企业。制造服务是向产品生产过程和产品使用过程所提供的各种形式服务的总称。前者为面向产品生产企业提供的各

种形式服务，如市场分析、产品研发、IT 服务、新工艺开发、制造资源维护、财务服务、人力资源开发等；后者为面向最终用户提供的各种形式服务，如产品运行服务、MRO、IT 服务、财务服务、技术培训、报废回收等。

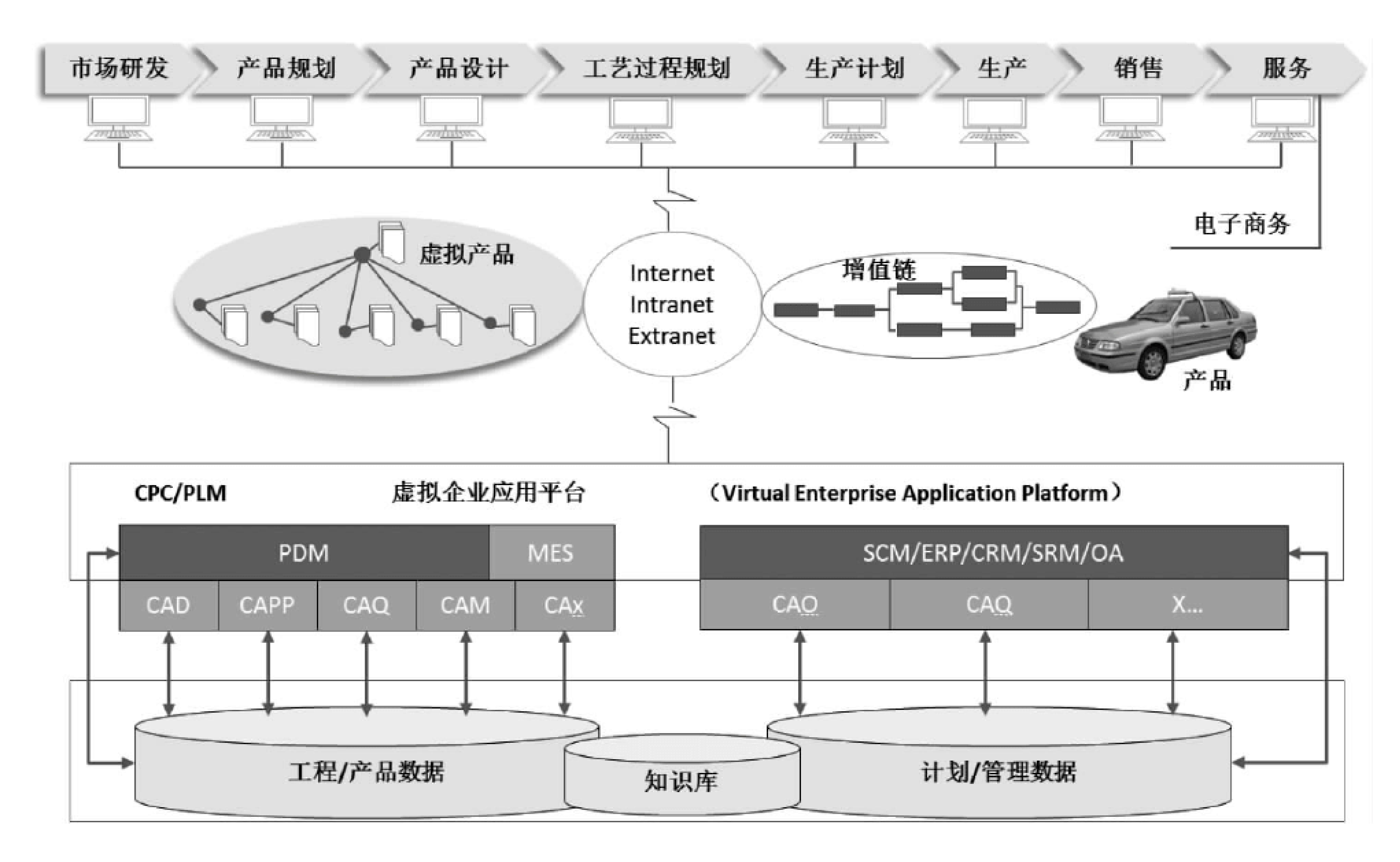


图 12-4　产品生命周期管理系统

12.2.3 集成和协同的重点和对象

从集成和协同的重点来看，已经从原先的信息集成、过程集成，发展到知识集成。目前的代表技术是知识管理和智能制造技术。

知识管理是指为提高企业竞争力而对企业知识的识别、收集、获取并充分发挥其作用的过程，其目标是使企业实现显性知识和隐性知识的共享，促进知识创新并最大限度地激发企业产品创新的核心要素，工业企业的发展逐渐从依靠资本积累转向依赖于知识积累与更新。各种显性知识和隐性知识将融入企业的产品、服务和生产过程，并作为产品进行生产，驱动以创新为目的的知识生产。而建立和挖掘客户的知识库和利用知识资源数据等作为最重要的知识管理系统的支撑技术将得到飞速发展。

从集成协同的对象来看，由于现代工程系统的复杂化趋势，包括工程系统的大型化、功能和结构的复杂化、追求目标的多元化等，以及多学科并行涉及的迫切需求，光、机、电、磁、液、信息等技术一体化趋势，使得现代工程系统的设计必须同时涉及众多不同学科或专业领域。在这种情况下，工程系统设计已经从单纯的几何模型、几何模型加部分性能模型发展到多学科模型，多学科设计优化方法应运而生。

多学科设计优化(Multidisciplinary Design Optimization，MDO)是一种用全局的观点，通过研究复杂工程系统与子系统之间的交互影响和协同作用，对复杂工程系统进行分析和优化设计的方法。实现多学科设计优化的技术和系统分别称为多学科设计优化技术和多学科设计优化系统。

MDO 的基本思想是：在复杂工程系统的设计过程中集成各个学科(或领域的知识)，应

用有效的设计/优化策略以及分布式计算机网络系统，组织和管理整个系统的设计过程，通过充分利用各个学科之间的相互作用所产生的协同效应，协调不同学科设计之间的耦合和可能出现的冲突，使复杂工程系统的设计从孤立的、串行的过程成为并行的、协同的过程，将设计的重点从单独的部件级转移到系统级整体性能优化。

多学科设计优化方法在复杂工程系统设计中的成功应用，使得设计人员能够在数值计算和仿真分析的基础上进行高水平的设计决策，大大提供了复杂工程系统的设计质量和设计效率，降低了开发成本。

12.2.4　主要的集成和协同技术

从主要的集成和协同技术来看，计算机技术和网络技术的发展，为制造系统的集成提供了很多有效的工具，使得原先十分复杂的集成工作变得非常简单。目前代表的技术如企业集成技术、物联网和云计算/云制造等。

企业集成从 20 世纪 80 年代到现在已经发展了将近三十年，从开始的点对点集成、企业应用集成、企业间集成，发展到了现在的面向服务的集成。随着信息技术的不断发展，集成的内涵不断发展，同时也促进了企业经营模式的变革。

物联网是通过射频识别、传感器网络、全球定位系统等信息传感设备，按约定的协议把任何物品与互联网连接起来，进行信息交换和通信，以实现智能化感知、监控和管理的一种网络。物联网是在互联网基础上延伸和扩展，其用户端延伸和扩展到了任何物品与物品之间进行信息交换和通信。物联网的特征有三点：接入对象更为广泛，获取信息更加丰富；网络可获得性更高，互联互通更为广泛；信息处理能力更强大，人类与周围世界的相处更为智慧。物联网的应用，将对工业信息化的发展产生巨大的影响。

随着网络基础设施的逐步完善，互联网、3G/4G/5G、无线宽带网络、无线传感等多种网络正在融合为泛在信息网络，“无时无刻不网络”的时代已经到来。在这种情况下，一种新的服务化计算模式——云计算(Cloud Computing)正在走向成熟。作为一种新的计算架构，云计算不仅对信息领域产生了重大影响，也对工业信息化的发展产生了重要的影响。云制造是借鉴云计算思想发展起来的一个新概念，是先进的信息技术、制造技术以及新兴物联网技术等交叉融合的产品，是“制造及服务”理念的体现。云制造需要采取包括云计算在内的当代信息技术前沿理念，建立共享制造资源的公共服务平台，将巨大的社会制造资源池连接在一起，提供各种制造服务，实现制造资源与服务的开发协作、社会资源高度共享。

12.3　中国制造信息化应用系统

工业制造主要应用的信息化系统可以分为工程设计自动化系统、制造控制自动化系统、柔性制造系统、制造执行信息系统、企业资源管理信息系统、信息物理系统等。本节将简要介绍这些系统功能。

12.3.1　工业设计自动化系统

工业设计自动化是指利用计算机软硬件及网络环境来辅助进行产品设计和分析的一种技术。即在网络和计算机辅助下，基于产品数据模型，对产品的设计、制造、装配、分析等过程提供计算机支持工具和手段。工程设计自动化不仅贯穿产品设计制造的全过程，而且涉

及企业的设备安装、物流配送、生产计划、成本控制等方面，其应用实施可以起到缩短产品研制周期、降低产品开发成本、实现产品优化设计的目的。工业设计自动化系统一般包括计算机辅助设计(CAD)、计算机辅助设计工程(CAE)、计算机辅助工艺设计(CAPP)、计算机辅助设计(CAE)、产品数据管理(PDM)等。

1. 计算机辅助设计系统

计算机辅助设计(Computer Aided Design,CAD)是指利用计算机系统辅助完成工程设计的产生、修改、分析、优化和检验的过程。CAD 技术从产生到现在，经历了形成、发展、提高和集成等阶段。在 CAD 技术发展的初期，CAD 仅限于二维计算辅助绘图，随着计算机软硬件技术的快速发展，CAD 技术从二维平面绘图发展到三维产品建模，之后产生了三维线框造型、曲面造型以及实体造型技术。现已向参数化及变量化设计思想和特征造型方向转变。

二维 CAD 系统将工程设计图纸看成是“点、线、圆、弧、文本”等几何元素的集合，所依赖的数据模型是纯几何模型，系统记录了这些图素的几何特征。二维 CAD 系统具有很强的交互式图形编辑功能，可以方便地对图形进行复制、删除和移动等操作，也包含尺寸标注、注解、形位公差标注、图形存储和管理等功能。三维实体模型具有二维绘图无法比拟的优点，例如，可以对重要零部件进行有限元分析与优化设计(CAE)，可以支持工艺规程(CAPP)生成和数控加工程序(CAM)，可以在模具制造之前利用快速成型的方法制造出装配检查及测试用的实物零件，也可以启动三维模型与二维图形的关联功能，自动生产二维工程图纸。

2. 计算机辅助工艺设计系统

工艺设计师产品制造过程中技术准备工作的一项重要内容，是产品设计与实际生产的纽带，是一个经验性很强而且随制造环境变化而多变的决策过程。工艺设计的任务在于：规定产品工艺过程、工艺操作内容、工艺装备(设备、工夹量辅具)和工艺参数等。常见的产品加工工艺包括：零件的机械加工工艺、钣金件的冲压工艺、零件的铸造工艺、锻造工艺、热处理工艺，以及装备工艺等。

计算机辅助工艺过程设计(Computer Aided Process Planning,CAPP)就是借助于计算机来制定产品的工艺规程、计算工艺参数、生成工序图，最终得到一份完整的加工工艺卡，并以此为依据进行产品的生产加工。CAPP 系统根据产品设计信息，首先完成零件信息描述；然后根据现有工艺人员的经验、标准工艺规范及工艺知识库中的信息，初步完成零件的工艺过程设计；再根据工厂装备、加工规则知识、设备的性能及加工精度，完成各工序、工步的设计；最后输出所要的工艺路线、工艺规程、材料定额、工时定额、工装明细表以及数控程序(NCP)等。

3. 计算机辅助制造系统

计算机辅助制造(Computer Aided Manufacturing,CAM)是指计算机产品制造方面有关应用的总和。广义上讲，CAM 是指利用计算机辅助产品制造过程中的直接和间接活动，包括 CAPP、NC 编程、工时定额的计算、生产计划的制订、资源需求计划制订等。狭义 CAM 是指与数控编程有关的内容，包括刀具轨迹规划、刀具文件生成、刀具轨迹仿真以及 NC 代码生成等。由于 CAPP、MRP、ERP 系统的发展，目前所提到的 CAM 大多是指狭义

CAM，即利用计算机辅助编制数控加工指令。它向上与CAD、CAPP实现无缝集成，向下方便、快捷、智能、高效地为数控生产服务。CAD中设计的结果(零件模型)，经过CAPP工艺编排产生工艺流程图后，最终在CAM中进行加工轨迹生成与仿真，产生数控加工代码，从而控制数控机床进行加工。

4. 计算机辅助工程

计算机辅助工程(Computer Aided Engineering，CAE)是用计算机辅助求解复杂工程和产品结构强度、刚度、屈曲稳定性、动力响应、热传导、三维多体接触、弹塑性等力学性能的分析计算以及结构性能的优化设计等问题的一种近似数值分析方法。CAE从20世纪60年代初在工程上开始应用到今天，已经历了五十多年的发展历史，其理论和算法都经历了从蓬勃发展到日趋成熟的过程，现已成为工程和产品结构分析中(如航空、航天、机械、土木结构等领域)必不可少的数值计算工具，同时也是分析连续力学各类问题的一种重要手段。随着计算机技术的普及和不断提高，CAE系统的功能和计算精度都有很大提高，各种基于产品数字建模的CAE系统应运而生，并已成为结构分析和结构优化的重要工具，同时也是计算机辅助4C系统(CAD/CAE/CAPP/CAM)的重要环节。CAE系统的核心思想是结构的离散化，即将实际结构离散为有限数目的规则单元组合体，实际结构的物理性能可以通过对离散体进行分析，得出满足工程精度的近似结果来替代对实际结构的分析，这样可以解决很多实际工程需要解决而理论分析又无法解决的复杂问题。其基本过程是将一个形状复杂的连续体的求解区域分解为有限的形状简单的子区域，即将一个连续体简化为由有限个单元组合的等效组合体；通过将连续体离散化，把求解连续体的场变量(应力、位移、压力和温度等)问题简化为求解有限的单元节点上的场变量值。此时得到的基本方程是一个代数方程组，而不是原来描述真实连续体场变量的微分方程组。求解后得到近似的数值解，其近似程度取决于所采用的单元类型、数量以及对单元的插值函数。

计算机辅助工程技术的提出就是要把工程(生产)的各个环节有机地组织起来，其关键就是将有关的信息集成，使其产生并存在于工程(产品)的整个生命周期。因此，CAE系统是一个包括相关人员、技术、经营管理及信息流和物流的有机集成且优化运行的复杂的系统。

5. 产品数据管理

产品数据管理(Product Data Management，PDM)可以看成是对工程数据管理、文档管理、产品信息管理、技术数据管理、图像管理及其他产品信息管理技术的一种概括与总称。最早出现于20世纪80年代初期，目的是解决大量工程图纸、技术文档以及CAD文件的电子化的管理问题，后来逐渐扩展到产品开发中的三个主要领域：设计图纸和电子文档的管理、物料清单管理以及工程文档的集成、工程变更请求/指令的跟踪与管理。由于PDM技术与应用范围发展很快，人们对它还没有一个统一的认识，给出的定义也不完全相同。从狭义上讲，PDM仅管理与工程设计相关领域内的信息；而从广义上讲，它可以覆盖到整个企业中从产品的市场需求、研究开发、产品设计、工程制造、销售到服务与维护等产品全生命周期中的信息。虽然PDM软件功能越来越丰富，但文档管理、工作流、项目管理、产品结构、配置管理与系统集成仍然是PDM系统的基本核心功能，目前企业实施PDM也主要集中在这些功能的实现上。PDM的基本原理是，在逻辑上将各个CAX信息化孤岛集成起来，利

用计算机系统控制整个产品的开发设计过程，通过逐步建立虚拟的产品模型，最终形成完整的产品描述、生产过程描述以及生产过程控制数据。技术信息系统和管理信息系统的有机集成，构成了支持整个产品形成过程的信息系统，同时也建立了 CIMS 的技术基础。通过建立虚拟的产品模型，PDM 系统可以有效、实时、完整地控制从产品规划到产品报废处理的整个产品生命周期中的各种复杂的数字化信息。

6. 产品生命周期管理

产品生命周期管理(Product Lifecycle Management，PLM)，按照 CIMDATA 的定义，主要包含三部分，即 CAX 软件(产品创新的工具类软件)、cPDM 软件(产品创新的管理类软件，包括 PDM 和在网上共享产品模型信息的协同软件等)和相关的咨询服务。实质上，PLM 与我国提出的 C4P(CAD/CAPP/CAM/CAE/PDM)，或者技术信息化基本上指的是同样的领域，即与产品创新有关的信息技术的总称。

从另一个角度而言，PLM 是一种理念，即对产品从创建到使用，到最终报废等全生命周期的产品数据信息进行管理的理念。在 PLM 理念产生之前，PDM 主要是针对产品研发过程的数据和过程的管理。而在 PLM 理念之下，PDM 的概念得到延伸，成为 cPDM，即基于协同的 PDM，可以实现研发部门、企业各相关部门，甚至企业间对产品数据的协同应用。

软件厂商推出的 PLM 软件是 PLM 第三个层次的概念。这些软件部分地覆盖了 CIMDATA 定义中 cPDM 应包含的功能，即不仅针对研发过程中的产品数据进行管理，同时也包括产品数据在生产、营销、采购、服务、维修等部门的应用。

因此，实质上 PLM 有三个层面的概念，即 PLM 领域、PLM 理念和 PLM 软件产品。而 PLM 软件的功能是 PDM 软件的扩展和延伸，PLM 软件的核心是 PDM 软件。

12.3.2 制造控制自动化系统

制造控制自动化系统是制造自动化分系统的硬件主体，主要包括专用自动化机床、组合机床、数控机床、加工中心、分布式数字控制(DNC)、柔性制造单元(FMC)、柔性制造系统(FMS)、柔性生产线(FML)等加工设备，以及测量设备、辅助设备(如刀具系统)、夹具装置等。还有传送带、有轨小车、自动导向小车、立体仓库、搬运机器人、托盘站等。

1. 数控系统

数控系统是指用数字量发出指令并实现产品加工与过程控制的系统，简称 NC (Numeric Control)系统。数控系统所控制的一般是位置、角度、速度等机械量，也有温度、压力、流量、颜色等物理量。这些量的大小不仅可用数字表示，而且是可测的。如果一台机床(如铣床、钻床、冲床、切割机床等)实现其自动工作的命令是以数字形式来描述的，则称其为数控机床。

2. CNC 系统

CNC(Computer Numerical Control，计算机数控)系统完成的功能与 NC 机床相同，只是 CNC 机床的逻辑控制、几何与工艺数据处理以及程序的执行都由一台(或多台)计算机完成，并且 CNC 处理的功能更为强大，增加了柔性。由于采用了计算机作为控制部件，CNC 系统通过常驻在计算机内部的数控软件实现部分或全部数控功能，从而能对机床运动进行实时控制。只要改变计算机的控制软件就能实现一种新的控制方式，这是 CNC 系统

的最大特点。整个 CNC 系统由计算机软硬件、输入输出设备、CNC 控制器、可编程逻辑控制器 PLC(大多内装在 CNC 控制器中)、主轴驱动单元和进给驱动单元等组成。

3. DNC 系统

DNC 是分布式数字控制(Distributed Numerical Control)或直接数字控制(Direct Numerical Control)的简称,是数控设备联网运行的基本方式。分布式数字控制除具有直接数字控制的功能外,还具有系统信息收集、系统状态监控以及系统控制等功能。DNC 中,有多台 NC、CNC 机床与过程计算机相连。过程计算机在大容量存储器中存取零件程序,并通过接口将这些程序传给各数控机床,完成 DNC 基本功能。

12.3.3 制造执行系统

制造执行系统(Manufacturing Execution System,MES)的概念最早形成于 20 世纪 80 年代末,20 世纪 90 年代后获得迅速发展。其目的是实现生产过程及其相关的人、物料、设备和在制品的全面集成,并对其进行有效管理、跟踪和控制,是最终实现制造过程的计划与物料流动、质量控制、工艺等的全面集成。

制造执行系统位于车间级并控制执行过程,具有十分重要的作用,它在计划管理层与底层控制之间架起了一座桥梁,填补了计划管理层和底层控制之间的“鸿沟”。MES 是面向车间生产过程的“实时”生产和调度,一方面 MES 可以将来自 ERP 软件的生产管理信息细化、分解,形成操作指令传递给底层控制;另一方面 MES 可以实时监控底层设备的运行状态,采集设备、仪表的状态数据,经过分析、计算与处理,触发新的事件,从而方便、可靠地将控制系统与信息系统联系在一起。

制造执行系统是面向制造过程的,它必然与其他的制造管理系统共享和交互信息,这些系统包括供应链管理、计划管理、销售和客户服务管理、产品及产品工艺管理、财务和成本管理以及底层生产控制管理等。图 12-5 反映了 MES 与企业其他管理系统之间的关系。

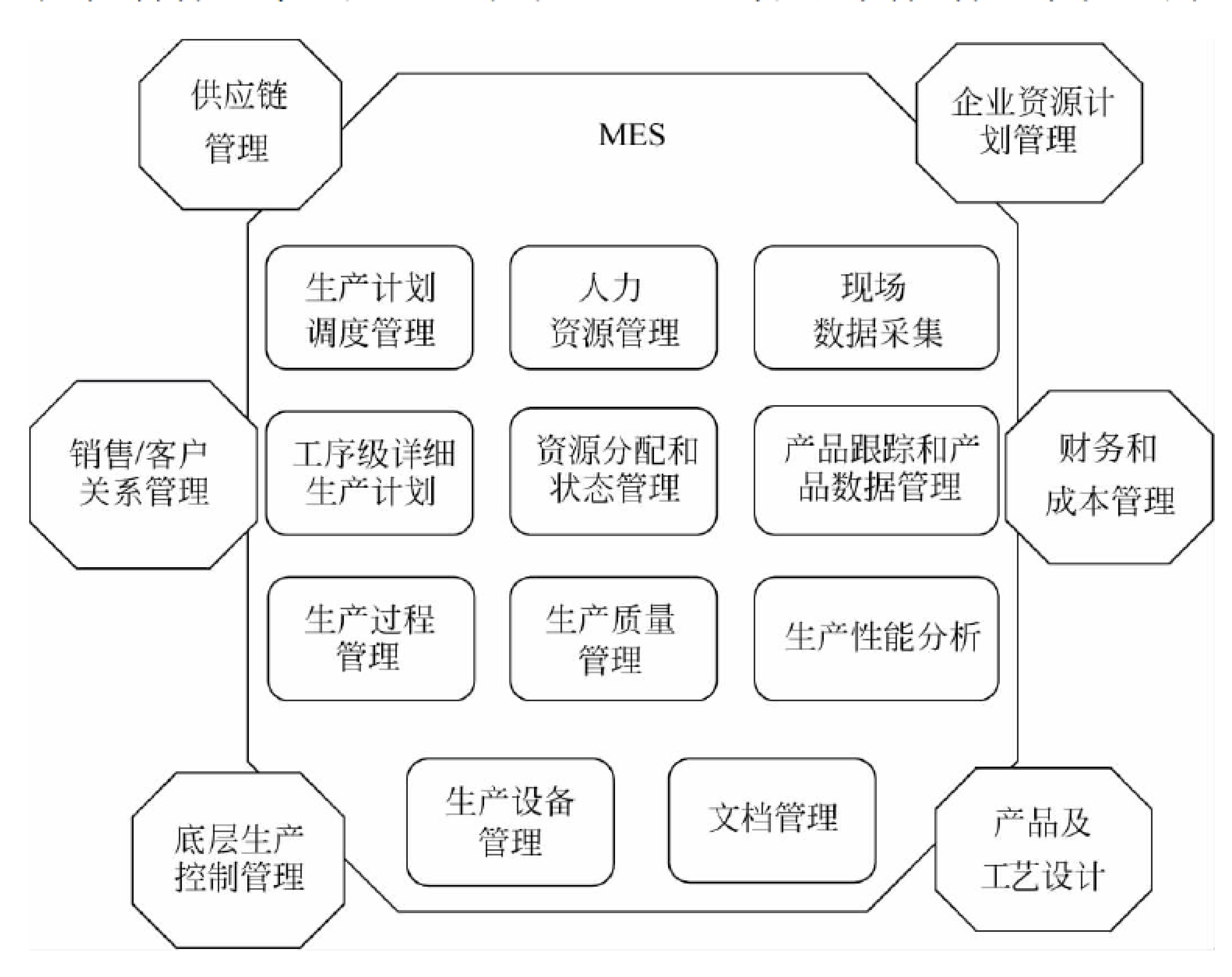


图 12-5　MES 功能模块图

MES 主要完成制造企业中的生产管理任务，根据国际 MES 协会 MESA 的定义，MES 系统的主要功能包括 11 个方面，如图 12-5 所示。这 11 个方面基本上囊括所有的生产管理要素，这些功能的取舍取决于特定的企业环境和期望的产出效益。

(1) 资源配置与状态跟踪。对所有的生产资料进行管理，包括机器、工具、劳工技能、材料等，使其井然有序，随时可以投入运转。同时记录资源的各种历史信息，以保证生产设备的配置，并对设备的实时状态信息进行跟踪。

(2) 工序/细节调度。根据生产单元的优先级、属性、特征，对生产工序进行优化，使生产资源配置的变化降到最低。

(3) 生产计划和调度。制定生产计划，并以任务、工单、批次、订单等形式下发给各生产单元。可以根据生产实绩实时调整原始计划，产生新的调度信息。

(4) 文档控制。对所有与生产单元有关的资料进行管理，包括工作指令、图纸、配方、标准操作流程、设计变更、产品记录以及 ISO 信息等，并进行历史数据的存储。

(5) 数据采集/获取。实时采集生产数据，记录生产单元的各种参数，并保存在相应的表格和记录中。数据可以由人工录入和从设备中自动采集。该功能需向外提供一个接口，以便其他应用可以通过它获得生产实时数据。

(6) 人力资源管理。记录员工的考勤以及专业技能。提供员工的实时状况记录，同时与资源配置功能交互，以产生最优配置。

(7) 质量管理。提供生产的实时分析，保证严格的产品质量控制。能够发现潜在的质量问题。对出现的质量问题进行诊断、分析，并提出改进方法。包括 SPC/SQC 的在线跟踪和离线分析功能。

(8) 过程管理：对生产过程进行监视，自动纠正或提示操作人员纠正生产中的纰漏。提供报警管理以及 MES 系统与智能设备的接口。

(9) 维护管理。跟踪并指导生产，以维护设备和工具。对突发问题做出快速响应。建立历史事件和故障记录数据库，协助完成故障诊断。

(10) 产品跟踪与记录。提供可视化的跟踪手段，监视产品的状态及用途。跟踪信息包括加工者、原料供应者、批号、序列号、当前产品状态、报警信息、返工及异常情况。通过跟踪信息可以追溯生产历史以及产品最终用途。

(11) 性能分析。对生产实绩和历史信息进行分析，以得到现实生产状况的效果，并与预计的效果进行比较。

12.3.4 柔性制造系统

柔性制造系统(Flexible Manufacturing System，FMS)是由统一的信息控制系统、物料储运系统和一组数字控制加工设备组成，能适应加工对象变换的自动化机械制造系统。FMS 的工艺基础是成组技术，它按照成组的加工对象确定工艺过程，选择相适应的数控加工设备和工件、工具等物料的储运系统，并由计算机进行控制。故能自动调整并实现一定范围内多种工件的成批高效生产，并能及时地改变产品以满足市场需求。FMS 兼有加工制造和部分生产管理两种功能，因此能综合地提高生产效益。FMS 的工艺范围正在不断扩大，包括毛坯制造、机械加工、装配和质量检验等。

柔性制造系统是一种技术复杂、高度自动化的系统，它将微电子学、计算机和系统工程

等技术有机地结合起来，理想和圆满地解决了机械制造高自动化与高柔性化之间的矛盾。它具有设备利用率高、生产能力相对稳定、产品质量高、运行灵活和产品应变能力大的优点。

FMS 可以分成以下 4 个等级：柔性制造模块、柔性制造单元、柔性自动线及柔性制造工厂。

(1) 柔性制造模块(Flexible Manufacturing Module，FMM)：是一台扩充了多种可任选功能(如刀具库、随行托架、交换装置等)的数控机床。

(2) 柔性制造单元(Flexible Manufacturing Cell，FMC)：一个 FMC 一般包括两三个 FMM，它们之间由工件自动输送设备连接。

(3) 柔性自动线(Flexible Tools Line，FTL)：又称柔性制造系统(Flexible Manufacturing System，FMS)。一般包括 4 台或更多台全自动 CNC 机床。各自备有搬运小车自动输送物料和一套计算机控制系统用以管理全部生产计划进度、物料搬运以及对机床群加工过程实现综合控制。

(4) 柔性制造工厂(Flexible Manufacturing Factory，FMF)：又称自动化工厂(Factory Automation，FA)。柔性由 FMS 覆盖到全厂范围，在全厂范围内实现生产管理过程、机械加工过程和物料储运过程的全面自动化，并由计算机系统进行综合控制。FMF 拥有分布式多级计算机系统(包括生产管理级主计算机)、自动仓库、十几乃至几十台各种 CNC 机床(加工中心、车削中心、CNC 车床、CNC 磨床、CNC 板材加工机床等)。FMF 是一种初级的 CIMS。

12.3.5 工业互联网与 CPS 系统

GE 的工业互联网与德国的工业 4.0 是应时代的技术基础和需求基础而产生的，均是基于当前的信息基础、市场需求、企业制造的成长及产品用户的体验而提出制造与服务的升级。工业互联网更多基于企业内部的制造升级和产品的运行与维护。德国工业 4.0 提出的 CPS 概念概括了工业制造的通用特征：物理设备与信息系统的协同。可以认为工业互联网和德国工业 4.0 提出的 CPS 是下一代工业制造的不同视角。

2012 年，美国 GE 公司提出将工业生产中的设备、数据和人进行有机结合，突破智慧和机器边缘，搭载互联网与工业连接，并称之为“工业互联网”。工业互联网的目标是通过机器和先进的传感器、控制和软件应用相连接，以提高生产效率、减少资源消耗。工业互联网的关键要素为：智能机器、工作人员、智慧分析。为推行工业互联网的理念，GE 提出了 1%的指标，并预测，每提高 1%的燃油效率，航空业每年能节省 20 亿美元，而能源行业则能节省 40 亿美元。

GE 通过自身制造体系实践工业互联网，在其产品中增加更多的传感器来获取海量数据，并最终帮客户提高其机车引擎、核磁共振仪器等设备的能源效率。在工业互联网战略下，GE 定位不再是软件公司和咨询公司，也不是装备公司，而定位自己是以资产为出发点，是一家服务公司，并通过智能机器的运营将数据服务作为自己最重要的产品。GE 的工业互联网将智能制造的制造环节和产品使用、运营维护连接在一起。

1. CPS 系统框架

“工业 4.0”是德国政府《高技术战略 2020》确定的十大未来项目之一，研究项目最初由德国联邦教研部与联邦经济技术部联手资助，在德国工程院、弗劳恩霍夫协会、西门子公司

等德国学术界和产业界的建议和推动下形成，德国政府在2013年4月的汉诺威工业博览会上正式推出“工业4.0”战略，其目的是为了提高德国工业的系统性竞争力。与德国类似，美国2012年提出《美国先进制造业国家战略计划》，旨在组建各领域的制造创新研究所(IMI)，从而建立起全国性的制造业领域的产学研虚拟联合网络。英国2013年提出《英国工业2050战略》。

“工业4.0”概念包含由集中式控制向分散式自组织的基本模式转变，目标是建立一个高度灵活的个性化和数字化的产品与服务的生产模式。“工业4.0”提出了信息物理系统(Cyber Physical System)，将制造业融合成为O2O形态。

工业4.0中的核心CPS平台即适应具有协作性特点的商业化进程和连接智能工厂和智能产品的全生命周期各方面的整个商业网络，因为其如下特点为企业的智能制造带来不仅技术变革更是商业模式的变化。

(1) 支持商业网络中互相协助的生产、服务、分析和预测；

(2) 适应具有协作性特点的商业化进程和连接智能工厂和智能产品的全生命周期各方面的整个商业网络；

(3) 提供迅速和简单流程的服务和应用；

(4) 在App Store模式链下实现商业进程中的调配和部署；

(5) 提供综合性强、安全可信的全商业进程支持；

(6) 保障从传感器到客户交流所有环节的安全和可靠系统；

(7) 支持移动端设备。

2. 工业4.0明确支持CPS平台的8大关键领域

(1) 标准化与参考架构。开发出一套单一的共同标准，合作伙伴关系才可能形成，需要一个参考架构为标准提供技术说明。

(2) 管理复杂系统的模型及相应方法。建立适应日益复杂系统的交互模型，提供开发这些模型所需的方法和工具。

(3) 基础设施。可靠、全面和高质量的通信网络是实现工业4.0的基础条件。

(4) 安全和保障。在通用安全标准下适应工业生产过程及产品数据信息的安全体系。

(5) 工作的组织和设计。智能工厂中员工的参与性工作设计及学习模型，工作流程、工作环境的重构。

(6) 培训和持续的专业发展。需要提供数字化学习计划及支撑数字化学习的相关技术体系。

(7) 监管框架。企业数据、责任问题、处理个人数据以及贸易限制等法规的适应性变化。

(8) 资源利用效率。需要实现在智能工厂中投入的额外资源与产生的节约潜力之间的平衡。

为实现工业4.0提出的三大主题：智慧工厂、智慧生产和智慧物流，策略文件提出了横向集成和纵向集成的两种方法。

横向集成是指将不同制造阶段和商业计划的IT系统集成在一起，既包括公司内部的材料、能源和信息的配置(例如，原材料物流，生产过程，产品外出物流，市场营销)，也包括不同公司间的配置(价值网络)，实现企业的价值网络。垂直集成是指将不同层面的IT系统

集成在一起(例如,执行器和传感器,控制,生产管理,制造和执行及企业计划等各种不同层面),实现网络化制造系统。横向集成和垂直集成最后实现贯穿整个价值链的端到端工程数字化集成。

高度动态配置与数据交换的标准化是工业 4.0 追求的两大目标。高度动态配置指机器系统通过网络实时获得相关信息,自主切换生产材料、生产方式,形成最佳配置;根据不同客户不同产品动态配置模块化生产线、模块化工厂。而数据交换标准化则包括工厂内部作业与生产线标准化、智能工厂生态链各环节标准化、制造业务应用系统之间的交换信息标准化。

在工业 4.0 战略文件中,创新制造业的商业模式是以解决顾客问题为核心。

12.3.6 ERP 信息系统

一般来说,企业常见的 ERP 功能模块有:供应链与客户关系管理模块、销售管理模块、产品设计管理模块、采购管理模块、计划管理模块、生产管理模块、库存管理模块、设备管理模块、质量检验管理模块、财务管理模块、人力资源管理模块。主要流程如图 12-6 所示。

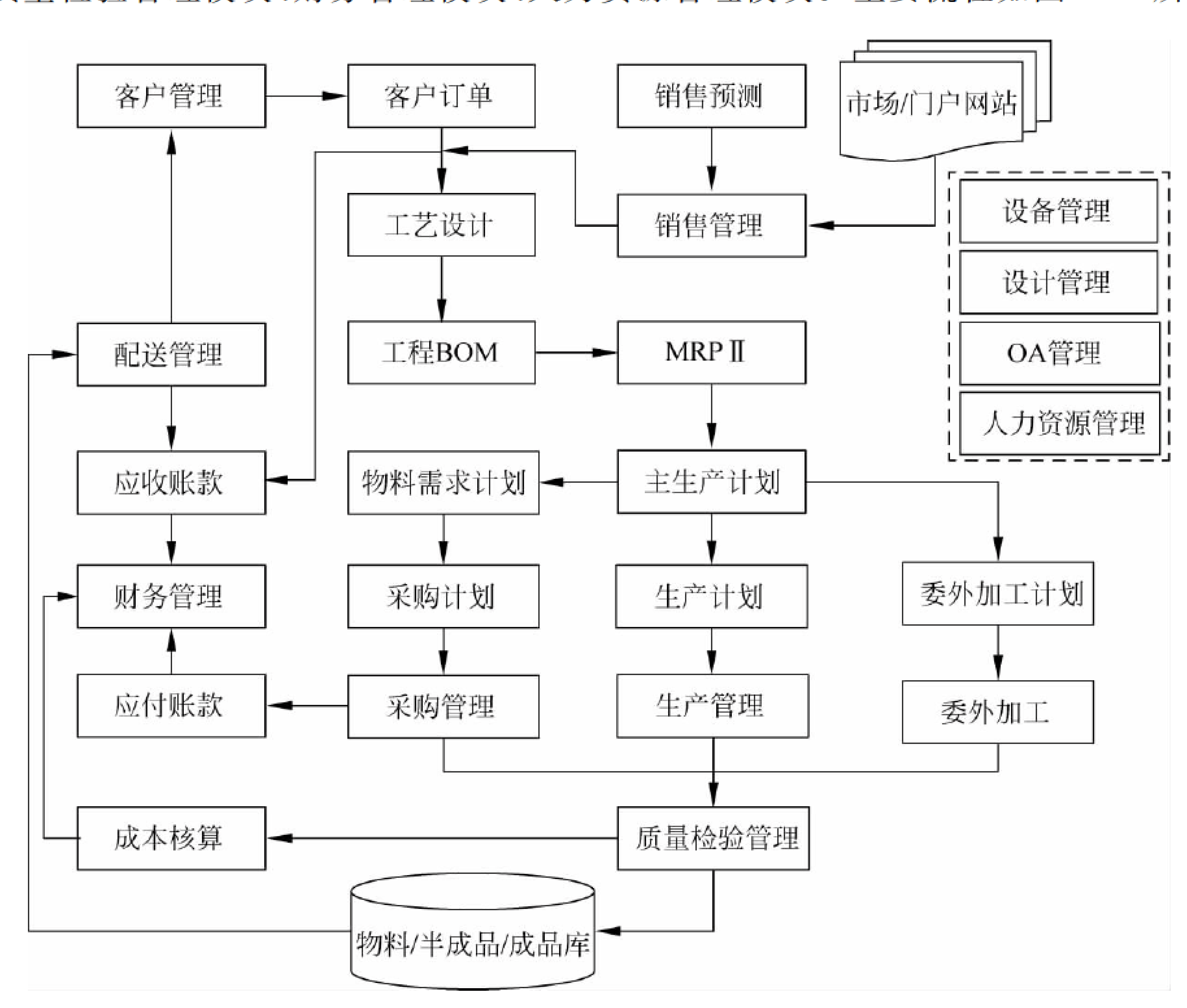


图 12-6 ERP 生产流程图

1. 供应链与客户关系管理模块

供应链管理(SCM)是对由供应商、制造商、分销商、零售商到顾客所构成的网络中的物流、信息流、资金流进行管理。供应链也称“需求链”或“价值链”,是实现最终顾客价值的综合过程。客户关系管理包括对客户档案信息维护;时间管理;潜在客户/项目管理/销售管理;合作伙伴关系管理;客户服务管理;市场/销售管理;客户档案维护/跟踪/分类;进行

销售预测；客户咨询、要求提供售后服务及反馈的受理；对客户进行满意度的分析。

2. 销售管理模块

销售管理是对销售合同、客户档案、销售出库、开票结账直到售后服务的销售业务全过程进行跟踪管理和统计分析，为制定营销策略提供决策依据。销售管理模块的功能主要有：制定销售计划和产品报价；根据相关信息制定销售订单；对销售合同进行管理；按照销售订单组织货源，安排发运，并将发货情况转交财务部；对销售情况进行统计、分析；开出销售发票，并向顾客催收货款。

3. 产品设计管理模块

产品设计管理是从明确设计任务开始，到完成图样和技术文件为止的技术工作过程。它包括使用要求的分析、设计方案（物流清单 BOM、产品档案和生产工艺）的优选、制图、试制、验证、形成图样和技术文件等内容。

4. 计划管理模块

计划管理的内容覆盖了企业各阶段所有职能的活动，其内容主要包括：主生产计划 MPS、物料需求计划 MRP、能力需求计划、采购计划、制造计划、委外加工计划等。

5. 采购管理模块

采购管理用来确定合理的订货量、优秀的供应商和保持最佳的安全储备。其功能主要有：对供应商进行管理；执行采购合同；能够随时提供订购、验收的信息，保证货物及时到达。

6. 生产管理模块

这一部分是 ERP 系统的核心所在，它将企业的整个生产过程有机地结合在一起，使得企业能够有效地降低库存，提高效率。同时使各个原本分散的生产流程自动连接，也使得生产流程能够前后连贯的进行，而不会出现生产脱节，耽误生产交货时间。它主要涉及：对车间作业进行管理和准时生产管理。

7. 库存管理模块

库存管理是企业记录、检查、跟踪、结存其库存活动的基础，是生产计划和库存控制系统中库存基础数据维护的主要环节。对物料管理，进行 ABC 分类分析，确定与采购决策相匹配的库存补充订货策略和订货批量计算方法。库存管理应当包括如下基本功能：仓库的发货和接收管理；保管、退货、盘点、调拨和预警管理；库存账务处理等。

8. 设备管理模块

设备管理使有限的设备资源，发挥最大的经济效益。其基本功能包括：建立设备、仪器、工装、模具和维修备件台账；编制设备维修计划；生产、测试设备运营管理；检查记录等。

9. 质量检验管理模块

ERP 中的质量检验管理主要对来料、在制品及成品进行质量检验。它的基本功能包括：制定检验计划；根据检验计划进行来料、在制品及成品检验；做好相关检验记录。

10. 财务管理模块

财务管理是 ERP 系统中的重要组成部分，它从货币的角度综合反映企业的生产经营情

况，通常财务管理可以由总账、固定资产、成本核算及控制、应收/应付账款管理、财务分析等构成财务管理的基本功能。具体如下：总账管理，包括建立科目体系、会计核算、多币种账务管理、现金管理及财务分析、自动产生财务报表等功能；固定资产管理，包括建立固定资产台账、固定资产折旧计算等功能；应收/应付账款管理，包括建立应收/应付账款，回款/付款、订金、退货/折让处理，应收/应付账月结转、坏账处理等功能。

11. 人力资源管理模块

以往的 ERP 系统基本上都是以生产制造及分销过程为中心的。因此，长期以来它一直把与制造资源相关的资源作为核心资源来进行管理。但近年来，企业内部的人力资源越来越受到关注，并被视为企业的资源之本。在此情况下，人力资源管理作为一个独立的模块，被加入到了 ERP 的系统中来，和 ERP 中的财务、生产系统组成了一个高效的、具有高度集成性的企业资源系统。它与传统方式下的人事管理有着根本的不同。现代人力资源管理主要包括：人力资源计划、招聘和选择、人力资源开发、报酬和福利、安全和健康、员工和劳动关系以及人力资源研究。

12.4　工业大数据架构体系

12.4.1　互联网催生工业大数据

工业 4.0 时代本质上仍然是企业互联网转型的一个重要部分和发展方向，它是生产制造过程的改进与变革，探索与互联网、物联网、大数据等融合基础上的工业革命，描绘不远的未来工业社会景象。就像德国工业 4.0 变革，德国和美国从不同角度给出了他们的答案，这也给中国制造产业升级给予很大启示。

德国工业 4.0 是在一个“智能、网络化的世界”里，物联网和务联网(服务互联网技术)将渗透到所有的关键领域，创造新价值的过程逐步发生改变，产业链分工将重组，传统的行业界限将消失，并会产生各种新的活动领域和合作形式。

美国在工业革命和互联网革命之后，2012 年 11 月 26 日，通用电气(以下简称 GE)发布白皮书《工业互联网：打破智慧与机器的边界》，提出工业互联网的概念。GE 的首席执行官伊梅尔特给出了所谓工业互联网的定义：

“开放、全球化的网络，将人、数据和机器连接起来。工业互联网的目标是升级那些关键的工业领域。”“这是一个庞大的物理世界，由机器、设备、集群和网络组成，能够在更深的层面和连接能力、大数据、数字分析相结合。这就是工业互联网革命。”GE 还大致描述了创新型工业互联网概念的理念，即通用平台、网络和数据的开放引入第三方创新者打造全新的服务和商业模式。GE 白皮书预测，在美国，如果工业互联网能够使生产率每年提高 1%～1.5%，使其重回互联网革命时期的峰值水平，那么未来 20 年，它将使平均收入比当前水平提供 25%～40%。

两者相比，美国工业关注的是生产过程的标准化和智能化，德国工业则是生产设备本身的智能化。美国工业努力减少人对生产过程的参与，提高生产线的柔性；德国工业努力提高设备安全性、降低能源损耗、降低设备维护量。二者从不同角度，充分利用互联网相关技术，对工业进行系统化、智能化的改造，最终结果殊途同归，实现一个万物互联的全新工业文明时代。

我国几乎所有的产业伴随着新一轮产业整合、劳动力成本的逐年上升、环境方面的恶化等因素，生产力涨幅有限的条件下，提出了供给侧改革。工业化必然要追求集约化、智能化、环保化等方面的变革和升级。

中国工程院院士、同济大学教授郭重庆在一次“互联网将重新定义制造业”的主题报告中指出，当今制造业价值链的每个环节——研发、设计、生产、销售、服务必须再定义，新的产品、新的流程和新的服务必须基于互联网的技术再造。企业的生存与发展将更多地依赖实施化市场洞彻，精确地满足消费者需求。互联网开源、开放、共创、共享的特性恰好能够从纵向供应链整合，到横向价值链整合上为制造业创造更多的发展空间。

“互联网＋工业”将成为未来制造业企业发展的范式(图 12-7)。“工业互联网将人和机器连接起来，将为制造商和客户带来前所未有的数据、信息和解决方案。”郭重庆认为，中国消费互联网企业基本上是在跟随和复制美国互联网企业的商业模式；而中国有偌大的制造业生产能力和消费市场，中国工业互联网完全可以跨越美国而抢先一步，为中国制造业的产业升级创造好的平台和机遇。



图 12-7　互联网＋工业(智能制造工厂)

12.4.2　工业大数据内涵特征

工业大数据是指在工业领域信息化应用中所产生的数据，是工业互联网的核心，是工业智能化发展的关键。工业大数据是基于网络互联和大数据技术，贯穿于工业的设计、工艺、生产、管理、服务等各个环节，使工业系统具备描述、诊断、预测、决策、控制等智能化功能的模式和结果。如图 12-8 所示为中国制造大数据分析。工业大数据从类型上主要分为现场设备数据、生产管理数据和外部数据。现场设备数据是来源于工业生产线设备、机器、产品等方面的数据，多由传感器、设备仪器仪表、工业控制系统进行采集产生，包括设备的运行数据、生产环境数据等。生产管理数据是指传统信息管理系统中产生的数据，如 SCM、CRM、ERP、MES 等。外部数据是指来源于工厂外部的数据，主要包括来自互联网的市场、环境、客户、政府、供应链等外部环境的信息和数据。

工业大数据具有 5 大特征。一是数据体量巨大，大量机器设备的高频数据和互联网数据持续涌入，大型工业企业的数据集将达到 PB 级甚至 EB 级别。二是数据分布广泛，分布于机器设备、工业产品、管理系统、互联网等各个环节。三是结构复杂，既有结构化和半结构化的传感数据，也有非结构化数据。四是数据处理速度需求多样，生产现场级要求实现实时

时间分析达到毫秒级，管理与决策应用需要支持交互式或批量数据分析。五是对数据分析的置信度要求较高，相关关系分析不足以支撑故障诊断、预测预警等工业应用，需要将物理模型与数据模型结合，追踪挖掘因果关系。

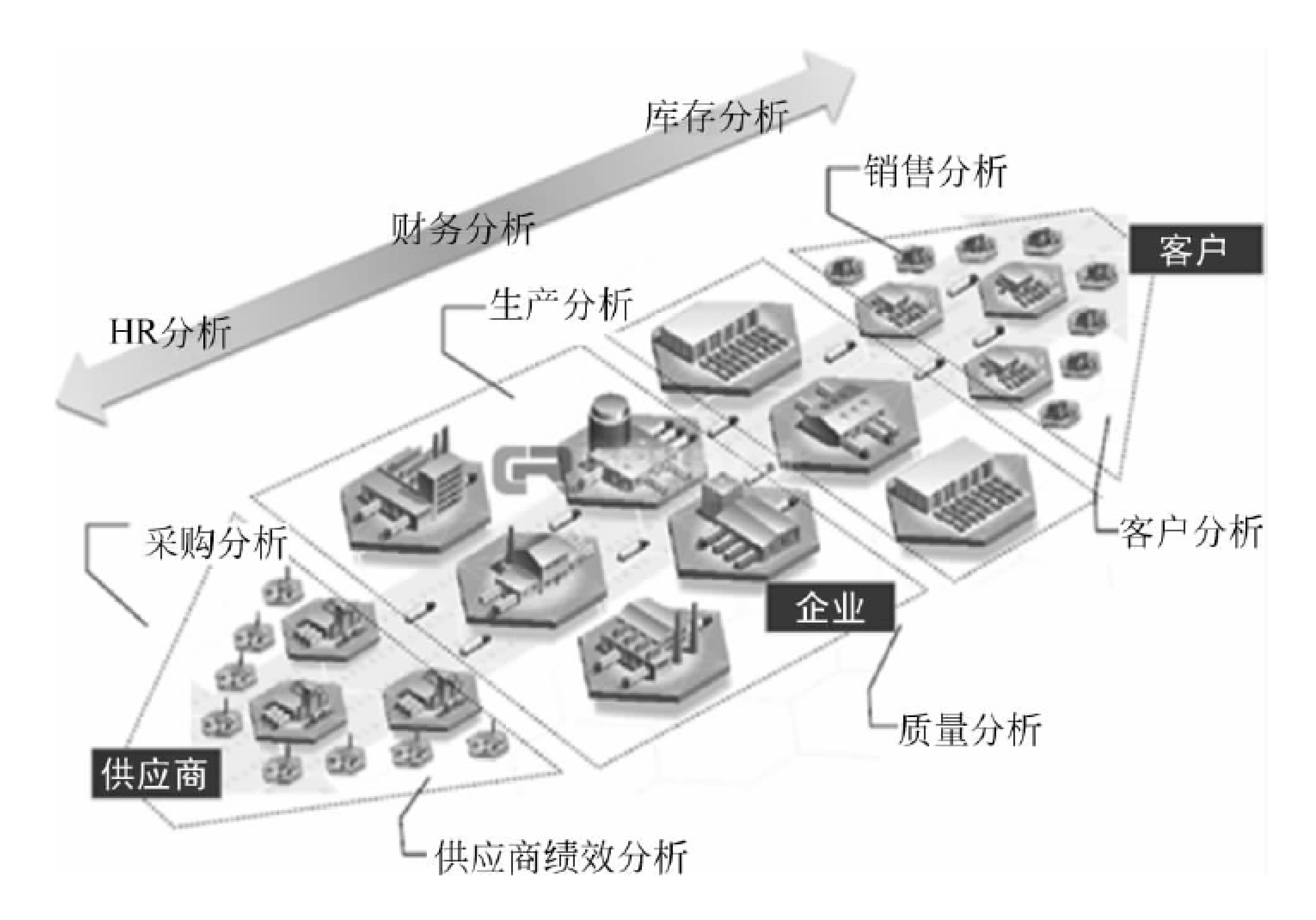


图 12-8　中国制造大数据分析

12.4.3　工业大数据业务架构

工业大数据的应用覆盖工业生产的全流程和产品的全生命周期。工业大数据的作用主要表现为状态描述、诊断分析、预测预警、辅助决策等方面，在智能化生产、网络化协同、个性化定制和服务化延伸 4 类场景下发挥着核心的驱动作用。工业大数据技术应用示意如图 12-9 所示。

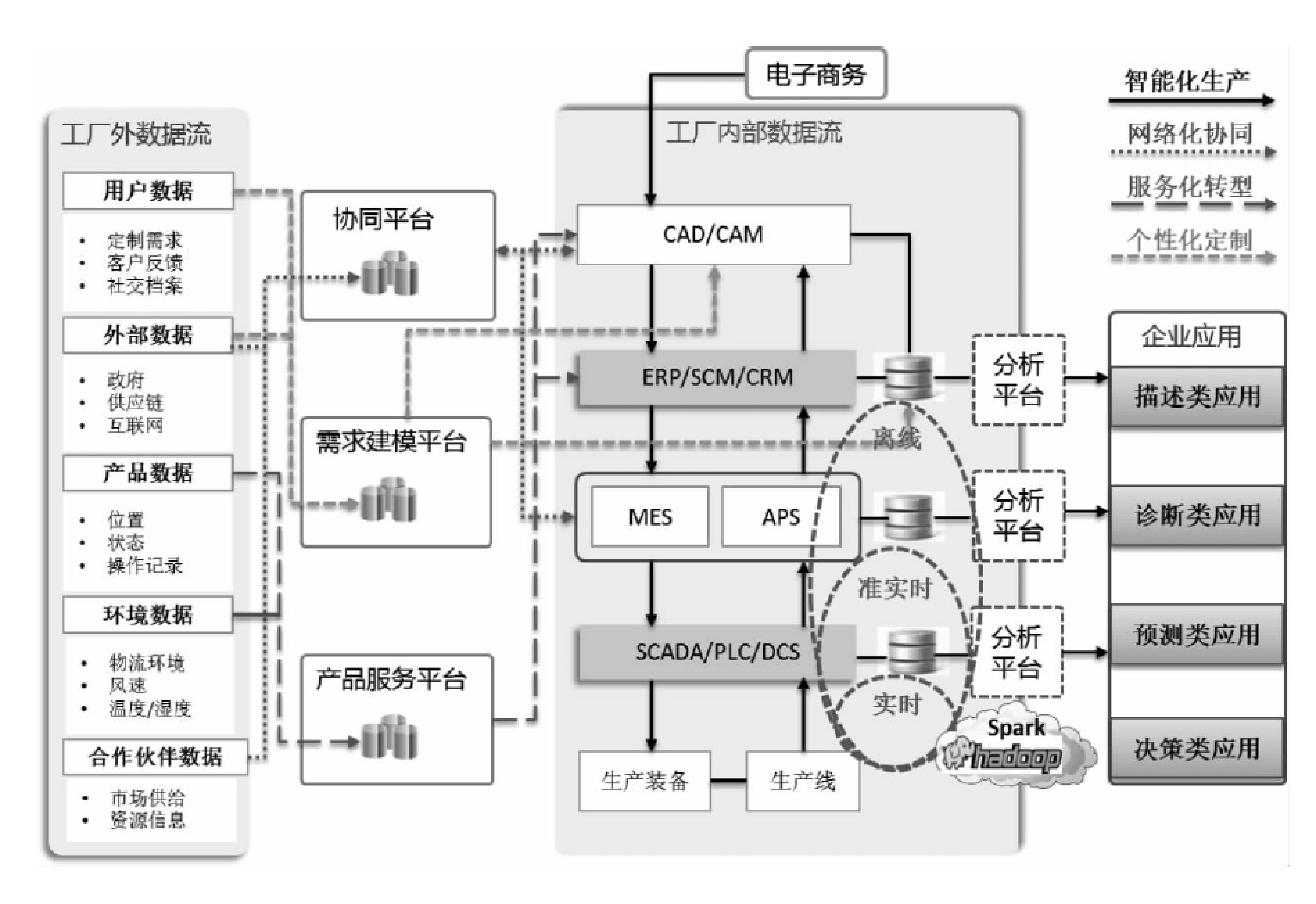


图 12-9　工业大数据应用示意图

1. 智能化生产中的工业大数据应用

虚拟设计与虚拟制造。虚拟设计与虚拟制造是指将大数据技术与CAD、CAE、CAM等设计工具相结合,深入了解历史工艺流程数据,找出产品方案、工艺流程、工厂布局与投入之间的模式和关系,对过去彼此孤立的各类数据进行汇总和分析,建立设计资源模型库、历史经验模型库,优化产品设计、工艺规划、工厂布局规划方案,并缩短产品研发周期。

生产工艺与流程优化。生产工艺与流程优化是指应用大数据分析功能,评估和改进当前操作工艺流程,对偏离标准工艺流程的情况进行报警,快速地发现错误或者瓶颈所在,实现生产过程中工艺流程的快速优化与调整。

设备预测维护。设备预测性维护是指建立大数据平台,从现场设备状态监测系统和实时数据库系统中获取设备振动、温度、压力、流量等数据,在大数据平台对数据进行存储管理,进一步通过构建基于规则的故障诊断、基于案例的故障诊断、设备状态劣化趋势预测、部件剩余寿命预测等模型,通过数据分析进行设备故障预测与诊断。

智能生产排程。智能生产排程是指收集客户订单、生产线、人员等数据,通过大数据技术发现历史预测与实际的偏差概率,考虑产能约束、人员技能约束、物料可用约束、工装模具约束,通过智能的优化算法,制定预计划排产,并监控计划与现场实际的偏差,动态地调整计划排产。

产品质量优化。产品质量优化是指通过收集生产线、产品等实时数据和历史数据,根据以往经验建立大数据模型,对质量缺陷产品的生产全过程进行回溯,快速甄别原因,改进生产问题,优化提升产品质量。

能源消耗管控。能源消耗管控是指对企业生产线各关键环节能耗排放和辅助传动输配环节的实时监控,收集生产线、关键环节能耗等相关数据,建立能耗仿真模型,进行多维度能耗模型仿真预测分析,获得生产线各环节的节能空间数据,协同操作智能优化负荷与能耗平衡,从而实现整体生产线柔性节能降耗减排,及时发现能耗的异常或峰值情况,实现生产过程中的能源消耗实时优化。

2. 网络化协同中的工业大数据应用

协同研发与制造。协同研发与制造主要是基于统一的设计平台和制造资源信息平台,集成设计工具库、模型库、知识库及制造企业生产能力信息,不同地域的企业或分支机构可以通过工业互联网网络访问设计平台获取相同的设计数据,也可获得同类制造企业闲置生产能力,实现多站点协同、多任务并行、多企业合作的异地协同设计与制造要求。

供应链配送体系优化。供应链配送体系优化主要是通过RFID等产品电子标识技术、物联网技术以及移动互联网技术获得供应商、库存、物流、生产、销售等完整产品供应链的大数据,利用这些数据进行分析,确定采购物料数量、运送时间等,实现供应链优化。

3. 个性化定制中的工业大数据应用

用户需求挖掘。用户需求挖掘主要指建立用户对商品需求的分析体系,挖掘用户深层次的需求,并建立科学的商品生产方案分析系统,结合用户需求与产品生产,形成满足消费者预期的各品类生产方案等,实现对市场的预知性判断。

个性化定制生产。个性化定制生产主要指采集客户个性化需求数据、工业企业生产数据、外部环境数据等信息,建立个性化产品模型,将产品方案、物料清单、工艺方案通过制造

执行系统快速传递给生产现场，进行产线调整和物料准备，快速生产出符合个性化需求的定制化产品。

4. 服务化延伸中的工业大数据应用

产品远程服务。产品远程服务是指通过搭建企业产品数据平台，围绕智能装备、智能家居、可穿戴设备、智能联网汽车等多类智能产品，采集产品数据，建立产品性能预测分析模型，提供智能产品的远程监测、诊断与运维服务，创造产品新的价值，实现制造企业的服务化转型。

12.4.4　工业大数据技术架构

工业互联网数据架构，从功能视角看，主要由数据采集与交换、数据预处理与存储、数据建模、数据分析和数据驱动下的决策与控制应用 4 个层次 5 大部分组成，如图 12-10 所示。

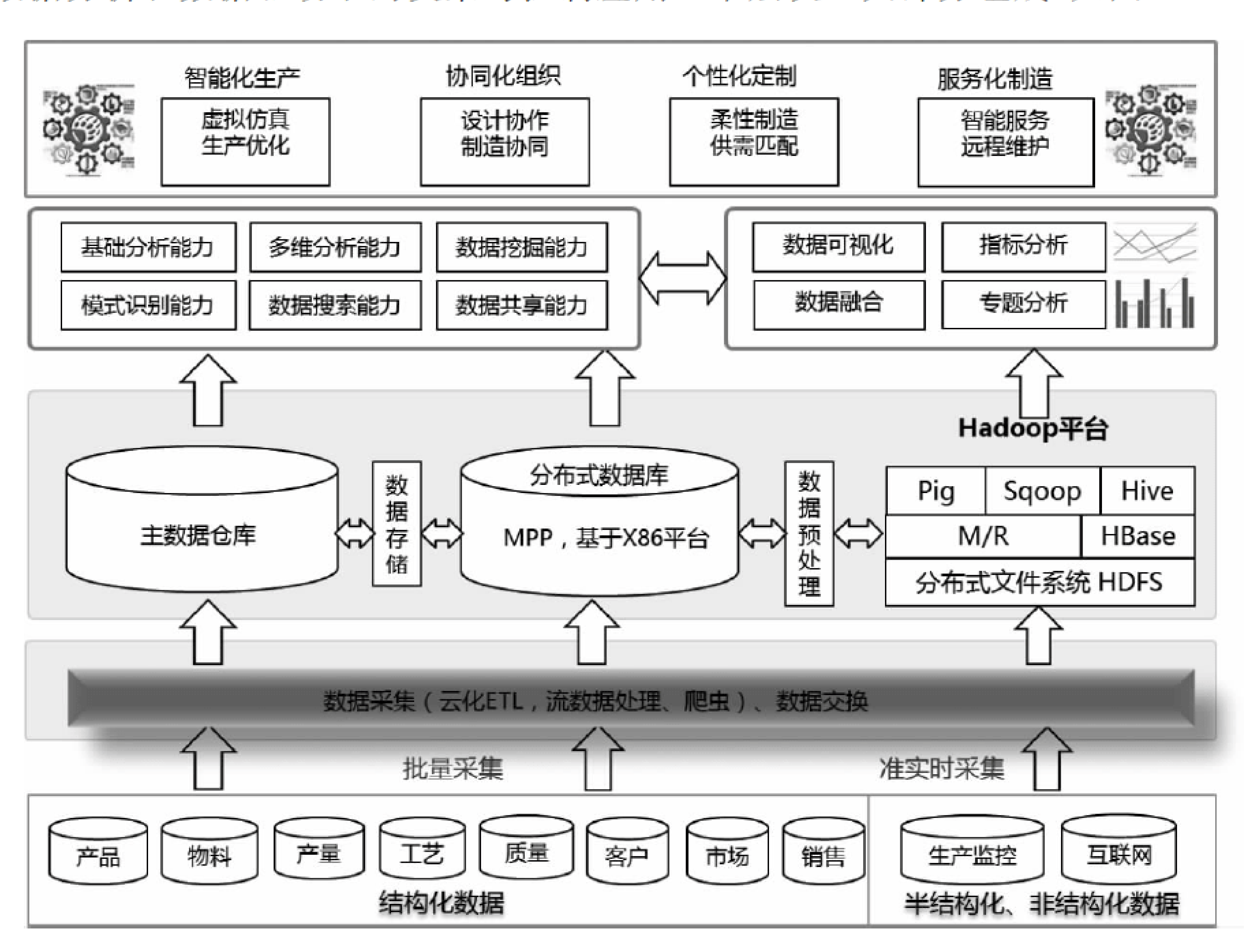


图 12-10　工业大数据技术架构

数据采集与交换层主要实现工业各环节数据的采集与交换，数据源既包含来自传感器、SCADA、MES、ERP 等内部系统的数据，也包含来自企业外部的数据，主要包含对象感知、实时采集与批量采集、数据核查、数据路由等功能。

数据预处理与存储层的关键目标是实现工业互联网数据的初步清洗、集成，并将工业系统与数据对象进行关联，主要包含数据预处理、数据存储等功能。

数据建模层根据工业实际元素与业务流程，在数据基础上构建用户、设备、产品、产线、工厂、工艺等数字化模型，并结合数据分析层提供数据报表、可视化、知识库、数据分析工具及数据开放功能，为各类决策分析提供支持。

决策与控制应用层主要是基于数据分析结果，生成描述、诊断、预测、决策、控制等不同应用，形成优化决策建议或产生直接控制指令，从而实现个性化定制、智能化生产、协同化组

织和服务化制造等创新模式，并将结果以数据化形式存储下来，最终构成从数据采集到设备、生产现场及企业运营管理持续优化闭环。

12.4.5 工业大数据安全架构

工业互联网的安全需求可从工业和互联网两个视角分析。从工业视角看，安全的重点是保障智能化生产的连续性、可靠性，关注智能装备、工业控制设备及系统的安全；从互联网视角看，安全主要保障个性化定制、网络化协同以及服务化延伸等工业互联网应用的安全运行以提供持续的服务能力，防止重要数据的泄漏，重点关注工业应用安全、网络安全、工业数据安全以及智能产品的服务安全。因此，从构建工业互联网大数据安全保障体系的数据安全与个人隐私等方面，包括支撑工业互联网业务运行的应用软件及平台的安全，工厂内部重要的生产管理数据、生产操作数据以及工厂外部数据（如用户数据）等各类数据的安全。说明如图 12-11 所示。

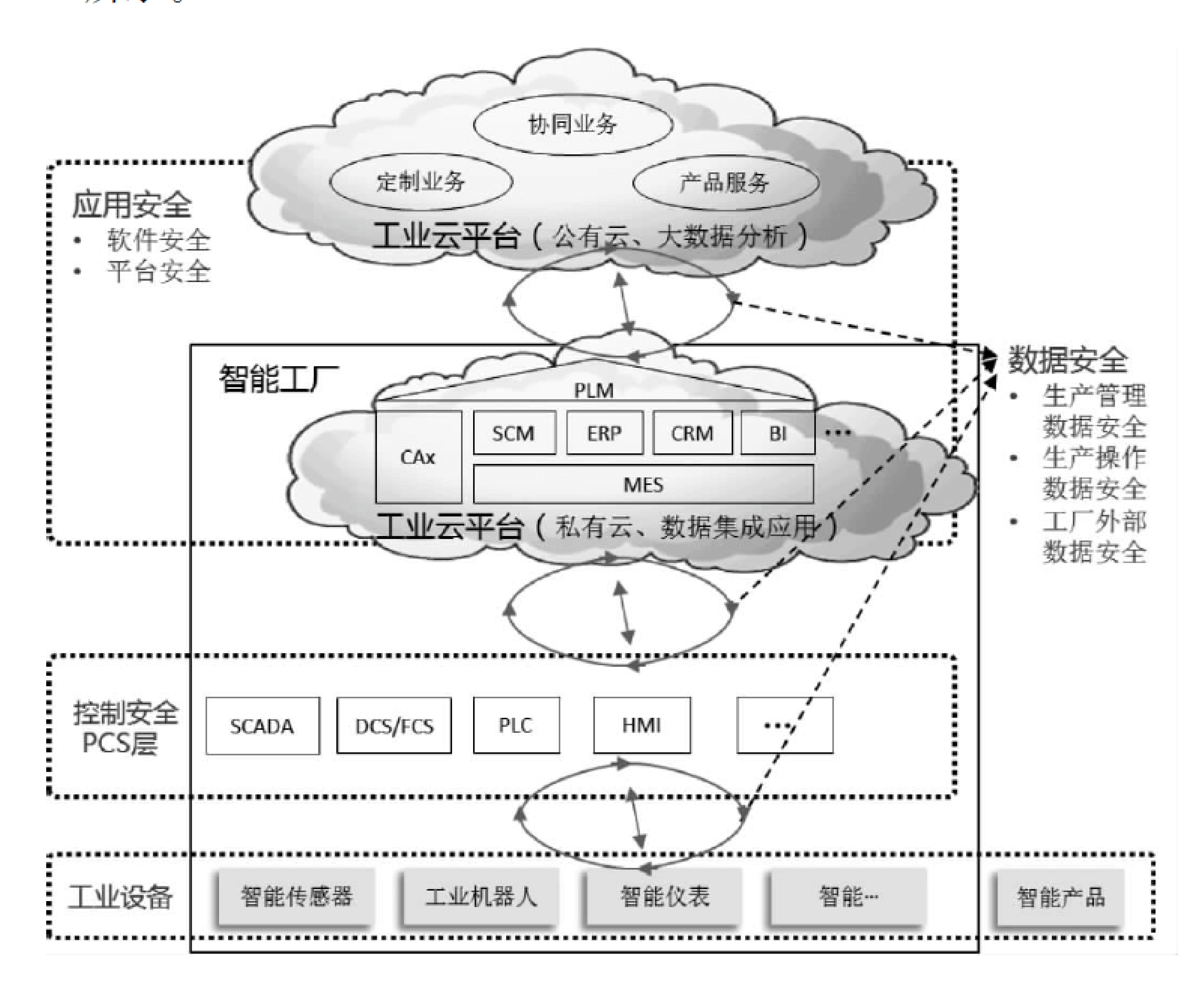


图 12-11 工业大数据安全架构

目前，工业领域安全防护采用分层分域的隔离和边界防护思路。工厂内网与工厂外网之间通常部署隔离和边界防护措施，采用防火墙、VPN、访问控制等边界防护措施保障工厂内网安全。从工厂内网来看，可进一步分为企业管理层和生产控制层。企业管理层主要包括企业管理相关的 ERP、CRM 等系统，与传统 IT 系统类似，主要关注信息安全的内容，采用权限管理、访问控制等传统信息系统安全防护措施，与生产控制层之间较多地采用工业防火墙、网闸等隔离设备，一般是通过白名单方式对工业协议如 OPC 等进行过滤，防止来自互联网的威胁渗透到生产过程。

在应用安全方面，网络化协同、服务化延伸、个性化定制等新模式新业态的出现对传统公共互联网的安全能力提出了更高要求。工业应用复杂，安全需求多样，因此对网络安全隔离能力、网络安全保障能力要求都将提高。并且将根据需要针对不同业务的安全需求提供

灵活的安全服务能力，提供统一灵活的认证、授权、审计等安全服务能力，同时支持百万级 VPN 隔离及用户量增长。在数据安全方面，工业数据由少量、单一、单向正在向大量、多维、双向转变，具体表现为工业互联网数据体量大、种类多、结构复杂，并在 IT 和 OT 层、工厂内外双向流动共享。工业领域业务应用复杂，数据种类和保护需求多样，数据流动方向和路径复杂，重要工业数据以及用户数据保护难度增大。需要采用工业数据以及用户数据分类分级保护机制。对重要工业数据以及用户数据进行分类分级，并采用不同的技术进行分级保护，通过数据标签、签名等技术实现对数据流动过程的监控审计，实现工业数据全生命周期的保护。

12.5　智能化协同制造体系架构

来自通信、网络、软件、自动控制等领域的技术进步推进了智能制造在微观技术与产品上的发展与应用，但是由于传统的智能制造主要关注制造的自动化、企业内不同业务系统的集成、生产线的柔性与工厂制造业务的敏捷性以及紧耦合企业集团内的协同，其协同机制在互联网环境下无法适应变化的企业联盟关系。工业互联网环境下协同制造更多表现为松耦合特征，因此自组织去中心化的企业间的制造服务趋于动态按需服务协同机制，基于智能化协同制造(ICM)体系结构也就出现了。ICM 与传统的智能制造的区别对应如表 12-1 所示。

表 12-1　智能制造概念及其特征比较

名　　称	含　　义	主 要 特 征	典 型 应 用	耦合性	动态配置
CIMS	计算机/现代集成制造系统	不同系统之间的集成与一体化	企业应用，企业联盟	紧耦合	弱
MAS	多智能体系统	系统或应用内部	企业应用	紧耦合	弱
AM	敏捷制造	企业间协作与集成	企业应用，企业联盟	紧耦合	弱
FM	柔性制造	生产设备单元之间的集成与协作	生产线，车间，工厂等企业应用	紧耦合	弱
Cloud-Manufacturing	云制造	应用服务的租用化与集中化	企业应用，企业联盟	紧耦合	弱
CPS	信息物理系统	产业链的横向与纵向集成	企业应用，企业联盟	松耦合	较强
ICM	智能化协同制造	以个性化客户需求为导向，以虚拟化资源平台为基础，制造服务化为转型，整合产业链上下游高效协作、高度协同制造	全制造服务生命周期的自组织和动态配置	松耦合	强

12.5.1　智能化协同制造发展需求

当今企业已经不再满足于规模的扩大，而越来越将其主要精力放在关注企业核心能力建设和核心竞争力的提升上。在工业互联网背景下，制造企业正在经历新一轮的大规模重组和优化，与过去企业内部组织优化方式最大的区别是，新一轮重组和优化是在整个产业链上展开的、面向全社会参与的新型协同生态系统(Collaborating Ecosystem)。智能化协同

制造发展需求可以从工业和互联网两个视角分析，如图 12-12 所示为智能化协同制造需求框架。

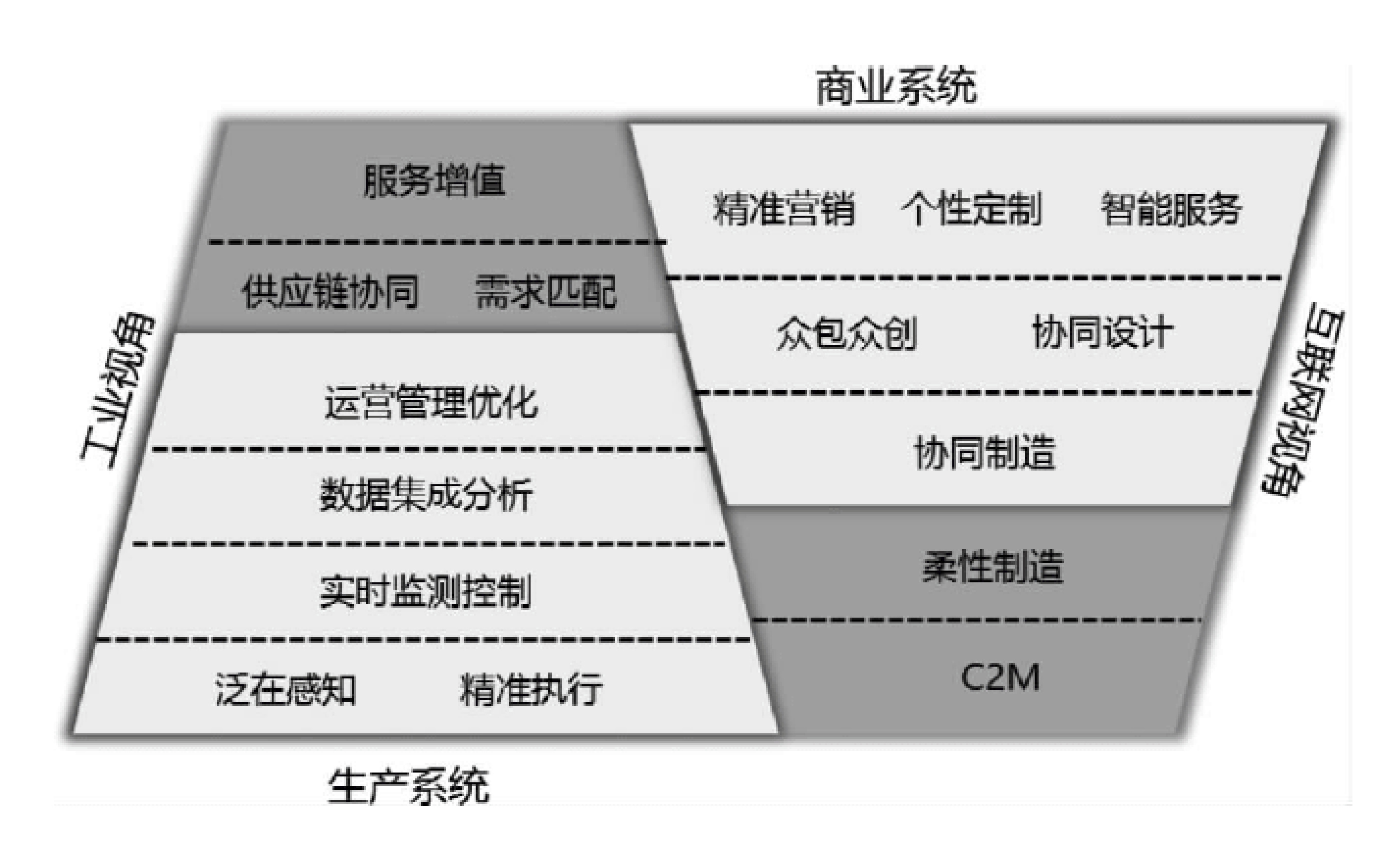


图 12-12　智能化协同制造需求框架

从工业视角看，智能化协同制造主要表现为从生产系统到商业系统的智能化，由内及外，生产系统自身通过采用信息通信技术，实现机器之间、机器与系统、企业上下游之间实时连接与智能交互，并带动商业活动优化。其业务需求包括面向工业体系各个层级的优化，如泛在感知、实时监测、精准控制、数据集成、运营优化、供应链协同、需求匹配、服务增值等业务需求。

从互联网视角看，智能化协同制造主要表现为商业系统变革牵引生产系统的智能化，由外及内，从营销、服务、设计环节的互联网新模式新业态带动生产组织和制造模式的智能化协同生态系统变革，其核心思想是以客户需求为导向的业务模式，其业务需求包括基于互联网平台实现的精准营销、个性定制、智能服务、众包众创、协同设计、协同制造、柔性制造等。

12.5.2　智能化协同制造总体架构

工业制造在网络互联、数据智能、安全保障等方面将进行快速的迭代演进，云计算和大数据技术逐步引入，扁平化的软硬件部署架构成为重要发展趋势，从而引发工业系统各层级网络、数据和安全的深刻变化。结合智能制造、互联网、数据、安全等发展趋势，智能化协同制造将随之产生。智能化协同制造目标架构如图 12-13 所示。

智能化协同制造目标实现架构主要呈现 4 个方面的关键特征。

1. 体系架构方面

实现层级打通、内外融合，传统工业系统多层结构逐渐演变为应用层、数据分析层和智能工厂资源层三层，整体架构呈现扁平化发展趋势。应用层按照智能化协同制造总体架构生命周期进行，包括产品前期市场研发、产品规划设计、工艺设计、生产计划、协同生产、协同销售和智能服务，这些业务的开展都已经突破传统意义的企业经营模式和发展理念，依据互联网虚拟生产线，这些制造资源能够按照流程任务完成和提交工作任务，像传统的企业协作模式一样，这样能够极大地发挥企业资源优势，最大限度地创造企业价值，也能够调动起员工的积极性、主动性和创造性。能够使跨地区、跨领域的员工为了实现同一目标，在分工明确的基础上彼此协作。而这种网络制造资源是计算机系统根据工作任务的要求，通过数据

挖掘优化选择和识别出来的,然后通过流程分配和调度工作任务。数据分析层完成工业大数据的数据集成、数据转化、数据预处理、数据挖掘和分布式数据存储等,提供应用层和智能工厂的协同工作。智能工厂(IM)是现代化制造企业的实体企业,比如海尔、格力等现代化制造车间,包括智能化制造企业具备的工业设备、控制系统以及ERP、MES、PLM等信息系统,还有机器人等。

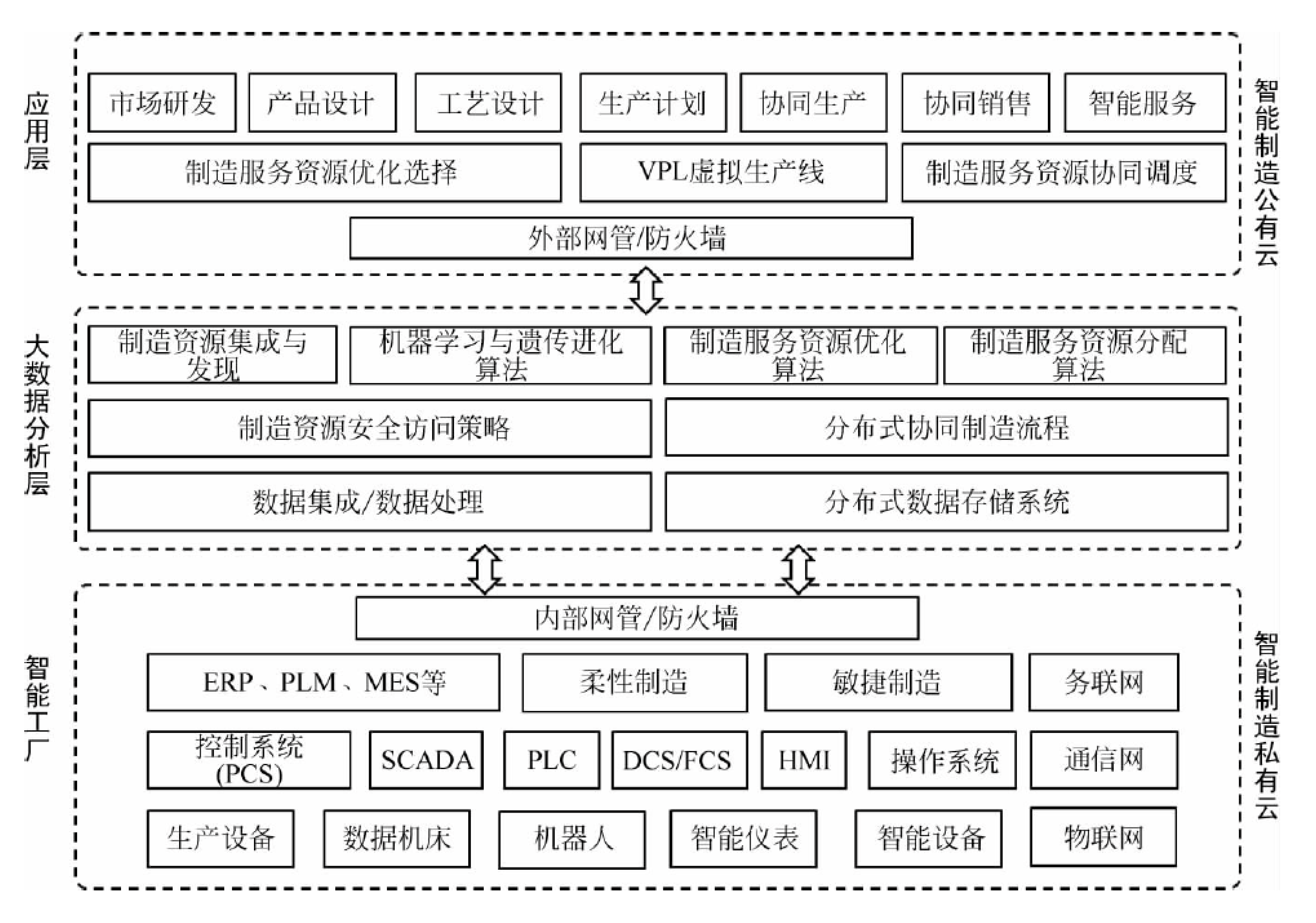


图12-13　智能化协同制造架构图

2. 网络互联方面

智能协同制造各种智能装备实现充分网络化,无线成为有线的重要补充,新型网关推动异构互联和协议转换,工厂与产品、外部信息系统和用户充分互联。智能工厂建立私有云、协同服务建立公有云,充分资源共享、互联互通,形成全面协同工作。

3. 大数据分析方面

工业云平台成为关键核心,实现工厂内外部数据的充分汇聚,支撑数据的存储、挖掘和分析,有效支撑工业信息控制系统和各种创新应用;涉及数据全面的采集与流动、工业数据云平台建设,以及多层次数据处理和分析能力构建,在此基础上支撑各种智能应用,同时应注意构建数据反馈闭环,以实现信息系统之间以及信息系统与物理系统之间的相互作用。这些数据主要包括工厂管理软件之间的信息交互,如研发设计类软件(CAD、CAE、CAPP、CAM等)、生产管理软件、ERP、客户管理软件CRM、供应链管理软件SCM等,实现这些管理软件之间的信息交互与集成;还有智能设备全面数据感知采集,包括采集机器、在制品等运行状态信息、采集生产环境信息、工业控制系统信息、机器人操作信息。这些数据利用云和大数据技术,推动智能工厂内部数据集成分析,同时构建决策反馈闭环,实现对工业生产的控制以及各种智能管理决策应用。通过工厂外部的工业云平台,汇聚产品数据、用户数

据、环境数据、协同企业数据等，并利用大数据技术，实现海量、复杂数据的综合存储、分析和处理。通过构建综合反馈闭环和评价体系，在工业云平台大数据集成与分析基础上，建立从工业云平台到企业级信息系统的综合性分析反馈闭环和客户评价，提升工厂内外的联动。

4. 安全保障方面

各种安全机制与工业互联网各个层次深度融合，实现纵深防御，立体防护，通过多种安全措施保障网络互联和数据集成安全。工业互联网目标架构的实现将是一个长期过程，需要网络、数据、安全等方面逐步协同推进。

12.5.3 智能化协同制造设计思想

智能化协同制造采用了面向服务企业、以服务互联创造价值的务联网（Internet of Service）以及企业服务总线的设计思想。

1. 面向服务企业定义

智能化协同制造是为了满足客户个性化定制的需求，企业业务流程和业务模式能够随需应变（On Demand Business）。随需应变可以理解为企业能够识别市场环境的变化，通过大数据分析预测能够预先洞察先机，先于其他竞争对手做出相应的调整和反应，保持客户、价值网伙伴和员工需求的同步。随需应变的业务能够带来业务整体柔性化、智能化和协同化。智能化协同制造的核心技术是业务组件化和面向服务。

业务组件是指业务组件给其他业务组件（内部或外部）提供的产品或服务。业务服务是业务组件的一个主要的特性。业务组件包括以下内容。

（1）业务目标：业务组件存在的理由，定义业务组件提供的基本价值。

（2）业务活动：业务组件内部执行活动的集合。

（3）业务资源：业务组件运作所需要的人、知识，任何有形或无形资产。

（4）管理机制：业务组件自治运作所需要的管理机制，包括对动机、性能和责任的评价指标和评价方法。

（5）业务服务：业务组件提供和消费的所有服务。

要实现随需应变的业务仅将业务组件化是不够的，企业的分解（或者称为业务的组件化）是将企业分解为一组更小的和自治的业务组件，这些业务组件在业务生态系统环境中与其他企业的类似组件进行交互，需要在整个价值网上实现业务组件间的无缝交互和紧密集成。同样，在整合价值网上实现业务柔性化要求组件网络必须具有柔性，即企业可以“内化（In-Sourcing）”外包得到的组件，或者“外包（Out-Sourcing）”其内部的组件。

面向服务是实现业务组件间无缝集成的核心，业务组件之间的交互体现了面向服务的思想，即每一个业务组件向其他业务组件提供一项或多项业务服务。使用业务组件服务的组件无须知道提供服务的业务组件是如何产生这个服务的。业务组件间的服务交互通过SLA（服务级别协议）来定义和约束，在 SLA 中定义了对交付服务的评价标准，用户根据 SLA 中定义了的业务层协议对服务进行管理。

在业务组件化和业务服务化的基础上，出现了所谓的“面向服务的企业（Service Oriented Enterprise，SOE）”的概念。面向服务的企业是一个通过 SOA 实施和对外发布其

业务流程的企业。通过将企业的业务单元组织成为提供各种服务的业务组件，在整个价值网络中，以服务提供和服务消费的方式实现企业内部不同业务单元（服务单元）和不同企业之间业务单元（服务单元）的业务协作，并按照事先约定的服务层协议对服务质量进行管理，快速柔性地响应市场需求的变化，实现企业和整个价值网络的利益最大化。

2. 业务组件建模

业务组件建模（CBM）是用来建立结构化业务组件模型的方法，它将企业的业务组件组织起来，在较高抽象层次上描述企业的业务逻辑，为解决业务低效问题和满足新的战略目标而实施的业务转型提供了基础。业务组件模型可以为企业提供业务战略和业务运作所需要的明确的重点领域和核心能力，可用来识别业务改进和创新的机会。通过重组企业当前的业务活动，形成一组可管理的、模块化的、可重用的组件，最终提高企业运作的柔性。

CIMS 的分析与设计涉及各类模型的建立，成熟的建模方法被广泛用于 CIMS 的实践中。ARIS（Architecture of Integrated Information System）是德国 Saarbrück 大学的 A. W. Scheer 教授于 1992 年提出的一种基于过程的模型结构。GRAI（Graph with Results and Activities Interrelated）方法由法国 Bordeaux 第一大学提出，专门为生产系统制定决策而开发的。20 世纪 80 年代初，美国空军 ICAM（Integrated Computer Aided Manufacturing）项目在 SADT（Structured Analysis and Design Technology）法的基础上发展了一套系统分析与设计方法，称之为 IDEF。它主要由 3 种模型组成：功能模型（IDEF0），信息模型（IDEF1X）和动态模型（IDEF2）。基于 BPMN 的建模方法借鉴了 UML 活动图、UML EDOC 的业务流程图、IDEF 等的技术经验，兼顾了复杂的流程语义和角色交互，为描述和研究复杂系统提供了手段。BPMN 由一组图形元素构成，便于开发一个简单的，为大多数业务分析人员熟悉的流程图。

业务组件建模是面向服务体系架构设计和业务流程管理的基础。图 12-14 给出了业务组件建模、业务流程管理、企业体系架构、面向服务的建模体系结构、面向服务的体系架构（SOA）运作、业务流程性能管理（BPPM）之间的关系。

3. 务联网概念

欧盟第七框架计划中提出的“未来互联网（Future Internet）”框架中，指出未来的互联网络架构由 4 个网络构成。人际网用于支持人-人之间的交流，如 Facebook 网站、博客、微信；物联网用于支持对物理世界运行状态和信息感知；知识和内容网支持知识的共享；务联网（Internet of Service）支持服务提供、服务组合和服务应用。欧盟研究人员认为务联网是关于未来互联网的一种观点，指的是所有需要使用软件应用的事务或事物都可以互联网上的服务形式存在，如软件、软件开发工具、软件运行平台等。

图 12-15 给出了一个务联网概念示意图。务联网以互联网和物联网作为手段，在现实的服务应用空间和数字化的虚拟空间之间建立联系，形成服务生态环境。与之相比，Internet 向用户单向发布信息，物联网从现实世界收集信息，云计算聚集资源并向顾客单向发布（计算的基础设施，通过特定的资源整合方式向客户提供各类网络计算资源）；它们可以被视为开环网络。务联网则通过“大规模定制”的方式为客户构建闭环网络：感知顾客大批量个性化服务需求，进而建立每个需求与可用服务之间的映射，面向服务功能/性能/价值等目标进行自适应的服务设计、选取与组合（计算资源、服务资源、社会资源），自适应地形成

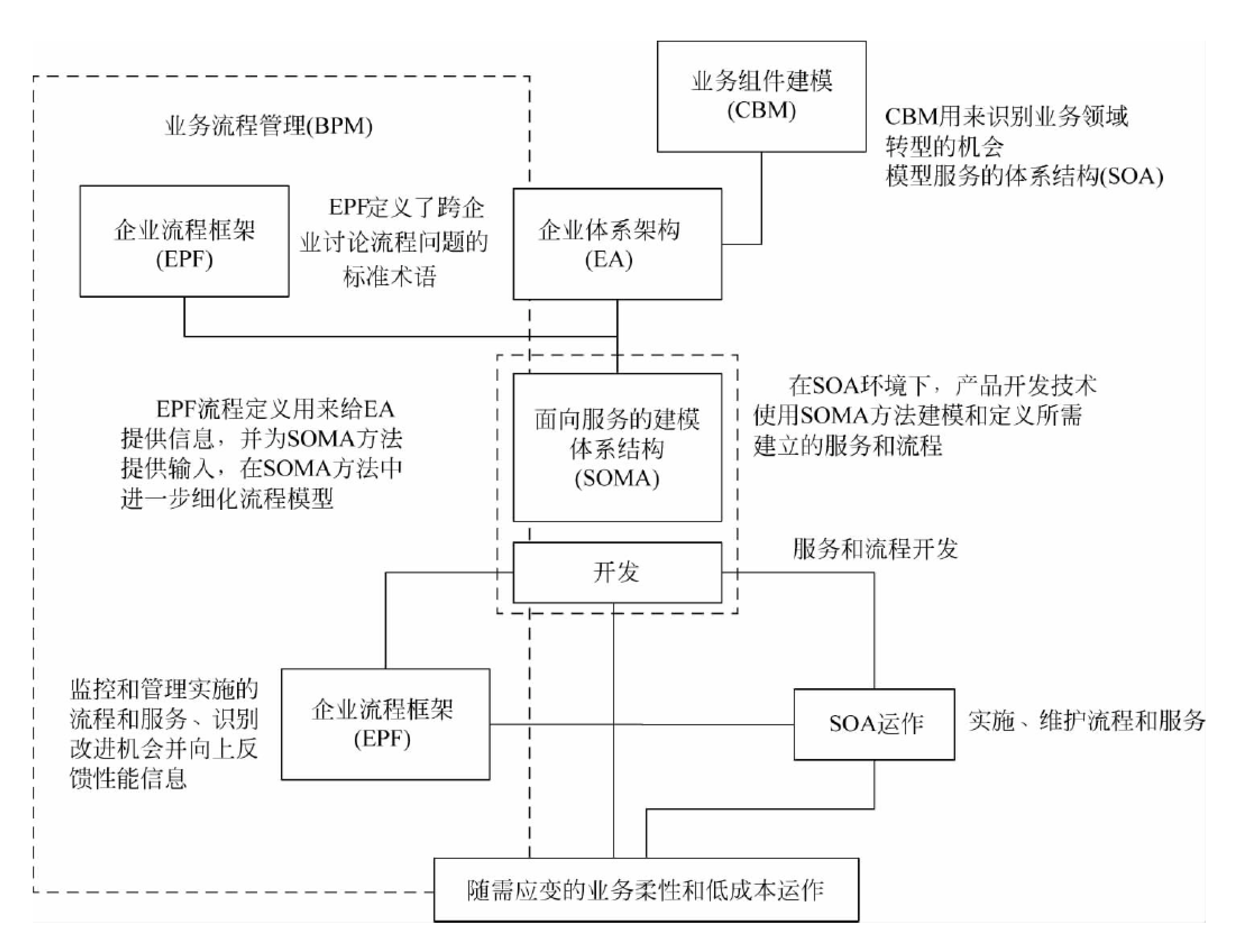


图 12-14　业务组件建模与 EA、SOA、BPM 的关系

务联网的子网络的集成，向每一个客户提供集成化服务，并根据应用情境的变化进行服务的演化；服务子网络中的各个服务节点在物联网和互联网的支持下进行协同，共同完成服务需求。

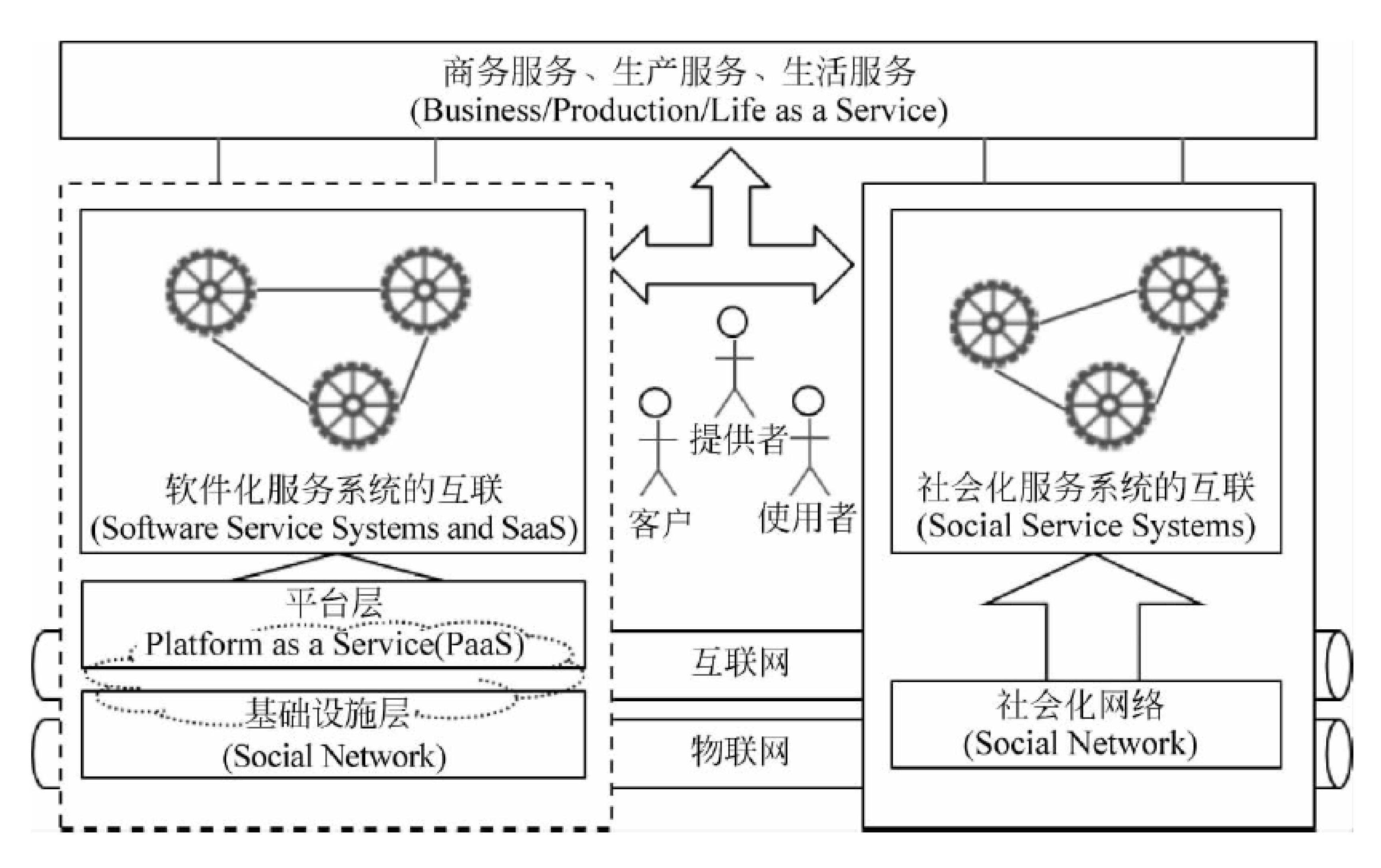


图 12-15　务联网概念示意图

4. 协同制造服务总线

为了实现智能化协同制造内若干个智能工厂之间的制造资源的即时装配、松度耦合、弹性可重构、自动集成、协同调用，需要构造协同制造总线(Collaborative Manufacturing Bus，CMB)。CMB 是智能工厂所有制造资源、制造服务应用之间交换数据以及智能工厂之间制造服务应用节点之间交换数据、智能工程与协作应用层的交换数据。CMB 是企业服务总线(ESB)的服务集成，也是一种资源协同调用的机制。可以想象全球企业是不可能实现所有的生产工艺、生产过程的完全一致的，但是可以要求企业都遵守微服务架构、基于 SOA 的注册、发布与使用，以及服务基于容器的运行。如图 12-16 所示为协同制造服务总线。

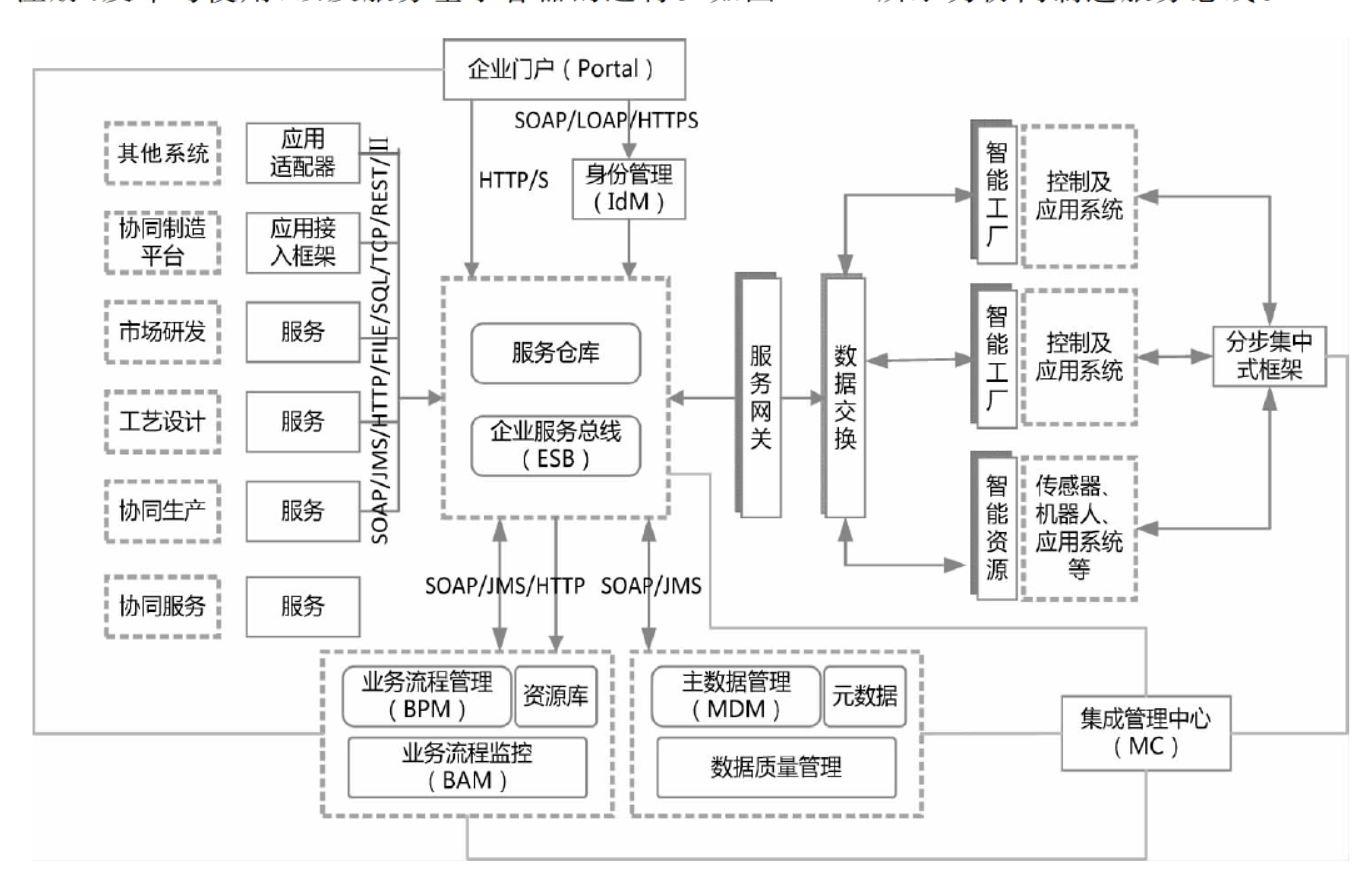


图 12-16 协同制造服务总线

考虑到 SOA 体系独立于硬件、操作系统和编程语言，CMB 原则上基于 SOA 设计原则，以 Web Service 方式来实现。各智能工厂的 CMB 通过通用的 SOA 机制中 WSDL 描述，UDDI 查找和发布，SOAP 来调用。

SOA 提供了跨平台跨语言松耦合的便利，当制造服务应用数量很少时，直接的点到点的 Web Service 服务接口是最快捷的集成方法。随着 IME 内服务应用的增多，单个服务应用的复杂性增强，功能增大，基本的 SOA 集成思维遇到挑战。在制造服务应用之间，即使基于 SOA 的 Web 服务，如果采用点对点的应用集成结构，尽管方法很简单，但存在着严重的隐患。用于连接的 Web 服务数目将快速增长(如果考虑方向性，总数为 $n\times(n-1)$，其中 n 为应用系统的个数)。不同应用系统之间由于缺乏自动提交请求的机制，必须在相关的连接 Web 服务内部固化请求的提交功能，应用系统之间存在着高度紧耦合，任何一个系统的升级或改动都将影响到其他与之相关的应用系统的修改。同时，当一个新的应用系统需要纳入整个应用集成体系时工作变得非常复杂。我们可以利用 Application Hub 来构造适合

于智能工厂内部应用协同或者紧耦合智能工厂之间应用系统的企业服务总线(Enterprise Service Bus,ESB)。智能工厂内部 ESB 通过服务交换点(Service Interactive Point,SIP)隔离单个 IME 内部的复杂 Web Service,形成一个抽象的 Web 服务集合。Web 服务端点、Web Sphere MQ 队列或 Java 的远程方法调用(Remote Method Invocation,RMI)远程对象的代理均可以提供 SIP 的实践方法。实现协议转换和消息透明路由与定位的中介弥补服务请求方与服务提供方之间协议最初的不一致。参与方(需求与提供者)均不需关心对方的位置或标识,本地的改变也自然不影响远程的参与方的活动,由此实现 Web Service 的虚拟化效果。

消息增强则弥补转换后的协议和被调用服务之间的参数差异,强化有效负载以确保调用规范化。转换则将消息从需求者的模式变为服务提供者的样式(如 SOAP/HTTP、JMS 和 MQ Integrator 等),期间会经过拆封、再封装、解密、再加密。相关聚合则从接收到的消息或者事件出发,根据需要派生或者触发必要的服务,以完成需求方的服务请求。我们可以通过设立一系列规则来完成模式标识和响应模式发现的行为,然后将得到的结果消息重组为源消息所需要的服务结果。

(1) 通过集成协同制造服务关键智能资源服务,并对这些服务进行有效的管理。

协同服务提供的关键业务能力可通过"服务化"进行规约,应用系统通过这些服务接口对外提供业务服务,需要通过集成对这些服务进行有效管理。

① 服务接口应在集成规划的基础上相对稳定;

② 服务需要资产化,得到统一的注册、管理和维护;

③ 服务的生命周期在"服务注册中心"得到统一的管理。

(2) 通过集成实现应用系统之间业务互通,使得应用系统之间业务易于协同。在"服务化"的基础上,打破应用系统之间的壁垒,使得应用系统之间业务流程实现互通,应用系统之间的协同以"服务调用"的形式进行,需要通过集成有效管理应用系统之间的服务调用交互,应用系统之间不应发生直接调用耦合,而是通过"协同制造服务总线"进行,使得服务交互双方能较快适应对方的变化。

(3) 通过集成规划新的协同应用,提升现有协同服务能力和水平。

在集成的基础上,通盘考虑,分析现有和未来业务需求,规划新的集成服务,提升业务水平。

协同制造服务总线内还需要建立的基本机制包括:元数据管理:在总线有效域内对服务的注册、命名及寻址进行管理。服务注册:从元数据中获取 SIP 描述、功能、与其他 SIP 的交互方式、QoS(Quality of Service)要求、语义注释等。

服务质量管理包括性能、服务的可交付以及如何对请求进行路由以实现负载均衡。QoS 策略可以封装在服务内,由需求方指定或者由提供方设置,也可以由 ESB 实现。在服务请求消息中,可以通过 QoS 参数来设置策略。事件监视观测消息从中介转换协议开始到服务交付完成是否发生异常,并记录日志。

服务管理针对如 JCA、Web 服务、Messaging、Adapter 之间的集成方式,对遗留系统适配器、服务编排和映射、协议转换方法、数据变换方法、企业应用集成中间件进行统一索引和定义。同时管理服务交互所需要的接口定义、消息模型,服务目录和发现等。

服务安全定义总线有效域内的认证和授权、服务交互的自动审核、数据安全标准的支

持、传输安全标准的支持。UDDI(Universal Description,Discovery and Integration)是基于 Web 的 Web Service 注册中心的实现标准规范,或者说是 Web Service 的目录服务,用于 Web Service 注册和发现。根据 UDDI 标准,注册中心可以部署为公共的、受保护的和私有的。

12.5.4 智能化协同制造应用场景

×集团化家电智能制造大型企业,要实现其全国各分公司协同制造,成员单位包含总装厂、设计研究所、配套企业,主要集中于武汉、上海、青岛、重庆、大连、昆明等城市。其中:

(1) 设计机构主要分布在北京、武汉和大连。

(2) 总装厂主要分布在青岛、大连、天津、山海关、武汉。

(3) 配套厂主要分布在太原、重庆、昆明、南京等地。

(4) 预研与咨询服务机构主要在北京。

由×集团成员单位地域分布特点,在产品的全制造过程中,需要异地协同并实现型号研制过程的全过程数据交换与管理。×集团针对数字化智能制造提出了如下总体要求。

(1) 预研服务需要在北京咨询中心和各企业的信息中心之间协同。各企业的信息中心在自己研究过程中积累的数据可以有条件分享给集团内其他企业。

(2) 异地设计师可以协同设计并进行文档修改并保持版本同步。

(3) 不同的分段可以分布式生产,最后实现总装。

(4) 车间实现刀具、机床等全覆盖管理并实现与企业内生产管理的集成。

上述目标均需要建立在×集团的大型单件式产品设计与制造的业务特点上。

12.6 智能化协同制造服务生命周期过程

工厂或者说企业个体是构造社会化智能制造体系的基础组织。我们将实现全在智能协同制造服务体系内,智能工厂之间基于互联网构成的去中心化企业网络,企业之间的联盟关系会随着制造任务的变化而变化。当企业内应用都基于 CMB 构造服务总线,且通过 UDDI 区域中心发布可分享的微制造服务组件时,智能工厂之间通过"资源发现-安全访问-资源选择-动态配置-共同进化"即可构成完整的协同制造链。

制造服务周期的智能工厂与传统的"工厂智能化"或"制造信息化"概念不同,企业的智能化是贯穿于全制造服务全生命周期。传统的"工厂智能化"或"制造信息化"聚焦于设计过程、制造计划管理、销售管理、物资管理、供应链管理、人力资源管理等,通常表现为 CAD、CAM、CAE、CAPP、ERP、SCM、MRP、HR、Portal 等系统以及系统间的整合。智能工厂的特点如下。

(1) 将智能化前置到市场研发。市场研发通常在产品设计的前端。当企业开展创新或者产品升级时,其如何升级的决策,来源于市场研发以及伴随的原型开发阶段。这一阶段以大数据方式进行广泛的数据收集整理和筛选、市场调研、竞争对手状况分析、结合产品的历史数据分析、前沿技术等进行数据挖掘和知识利用,以准确预测发展趋势。

(2) 将智能化后延到产品交付后的产品运行、维护、客户服务和状态监控的智能服务中。与传统的产品交付即制造周期的完结不同,产品交付后用户使用过程中的状态参数的

收集、用户评价的反馈、用户社区运营后积累的信息、用户体验的反馈、售后服务过程中维修数据的积累应以一定的模型反馈到产品的更新与升级状态中。

(3) 智能工厂还需要使得客户成为制造环节中的重要元素,而不仅仅是企业内部的管理者、工程师、技师和服务人员。客户有明确的系统入口来主动决定个性化产品的参数和属性,并直接以特定的身份参与到智能工厂整体系统的运行。由此,与传统的制造过程有所区别,我们定义全制造服务周期如下。

全制造服务周期(Total Manufacturing Service Lifecycle,TMSL)指包括产品从创新到运行,具体指由企业内包括市场研发、创新决策、开发与设计、工艺与制造、生产计划与管理、物流与供应链、营销交付、运行维护、用户服务过程形成的封闭信息环路牵引的全部制造过程的集合。

智能化协同制造是指在整个互联网中,智能工厂间,或者智能工厂与消费者之间发生的一种联合机制。该机制实现基于点到点、自组织的智能制造资源目录分享、同步;微制造服务组件发布、搜索、调用。基于制造资源和微制造服务组件建立在全网动态配置的虚拟生产线的建模和驱动。安全机制包括针对智能制造资源服务和制造服务单元组件节点的认证与授权机制、访问控制列表、数据传输 SSL 与 TSL、数据内容加密以及其他安全防护手段。还有比如企业的设计工作主要依赖于 CAD 平台。设计平台从单机到网络,然后进化到基于云服务器和虚拟化的 CAD SaaS 服务。基于 SaaS 服务的云设计机制主要由多人针对同样一个任务进行异步工作,以确保最终设计文档的同步。与此不同,协同设计服务不仅包括上述基本功能,更重要的是,无论是个体还是协作企业,通过自主设计任务或者按需设计任务的不断微型分割,微型设计组件作为可租售的成果在协同设计仓库中被全部设计需求者可见,即协同设计成果的商品交换活动就会自然发生。传统的工艺管理则是 CAPP 系统。同样目前 CAPP 多是网络版本,运行在企业内部。少量 CAPP 正迁移到云 CAPP SaaS 系统上,以保证工艺过程文档的版本管理和一致性。企业可以租用 CAPP SaaS 以降低硬件、软件和平台建设投入,也不需关心 CAPP 升级带来的维护问题。但是协同工艺不仅提供上述功能,还可以就某个零部件加工的工艺数据进行智能协同制造体系内的询问、调用。传统服务最初只是限于企业内部设备维修维护、操作管理。扩展的智能服务延伸到企业产品的售后运行维护跟踪与客户服务、产品用户体验管理与数据分析。同时提供协作各地的技术专家人力资源构成面向最终客户的智能服务,距离优先选择可以设定为,挑选距离客户服务请求地最近的服务伙伴或者服务专业人员去响应;技术可靠性优先选择我们则设定为根据智能服务应用之间的数据分析,选择最适合、最具有经验解决问题的技术专家去提供服务响应。

12.6.1 制造资源服务集成与发现

1. 资源服务引擎

制造服务注册中心为网络化敏捷制造平台提供了一个良好的制造服务发布、维护和管理环境,是构建协同制造链的基础。在实现对制造服务进行基于语义的描述以及匹配时,都需要这样一个稳定和可靠的注册中心作为支撑。作为制造服务注册中心的关键模块之一,制造服务匹配引擎则是整个协同制造链构建支持系统的核心,其功能主要是实现协同制造任务与制造服务注册中心中制造服务的匹配计算,并基于用户定义的匹配度以发现满足要

求的制造服务，本书提出的制造服务匹配引擎结构图如图 12-17 所示。

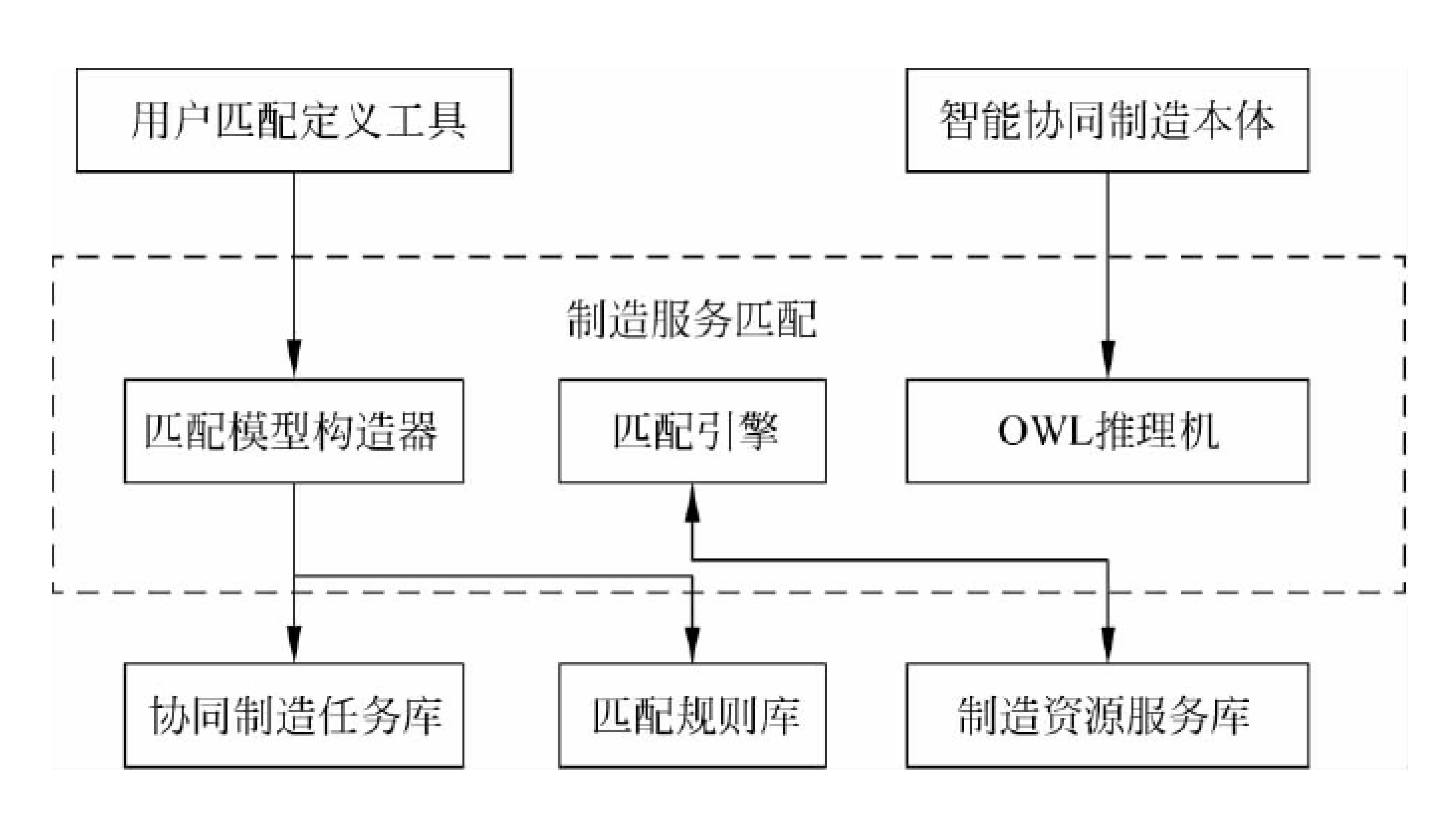


图 12-17 制造服务匹配引擎结构图

由图 12-17 可知，制造服务匹配引擎主要由匹配模型构造器、匹配引擎以及 OWL 推理机组成。制造服务匹配引擎的具体工作流程如下。

(1) 协同制造链发起企业首先使用用户匹配定义工具设定协同制造任务与制造服务的匹配度，不同匹配度等级的设定将直接影响到匹配结果的精确性。进行制造服务匹配时，匹配引擎将依据用户设定的匹配度进行匹配，发现合适的制造服务。用户还可以通过用户匹配定义工具设定特殊的制造服务匹配要求，如有关制造服务质量、制造服务提供商所在地理位置等协同制造任务制造能力特征所无法反映的要求。

(2) 匹配模型构造器从协同制造任务库中读取需要进行匹配的协同制造任务 OWL 描述文档，并结合用户匹配定义工具定义的匹配要求，生成该协同制造任务的语义匹配模型。在具体生成匹配模型的过程中，匹配模型构造器还需要访问匹配规则库中相关匹配规则信息。由于本文所建立的协同制造任务语义描述模型没有包含数值类型数据的比较关系，故本文预先建立了很多匹配规则，并将其存储于匹配规则库中。如对于零件尺寸，面向协同制造任务，系统建立了大于等于的规则，即只要制造服务的尺寸能力大于等于该零件的轮廓尺寸，其尺寸能力可视为满足要求。

(3) 匹配引擎从制造服务库中读入制造服务 OWL 描述文档，并通过 OWL 推理机将制造服务 OWL 描述文档与网络协同制造本体绑定，依据 OWL 语义逻辑生成制造服务推理模型。制造服务推理模型是基于网络协同制造本体，包含各种概念、属性及其之间扩展关系的模型。

(4) 匹配引擎将协同制造任务的匹配模型与制造服务的推理模型进行匹配度计算，获得满足用户设定的匹配度要求的制造服务。

2. 制造资源服务匹配数据挖掘

严格意义上来说，当制造服务的制造能力特征与协同制造任务的制造能力需求完全一致时，称制造服务在能力上完全满足协同制造任务的要求。显然，这种定义过于严格，因为制造服务提供商与协同制造链发起企业事先不可能就制造服务与协同制造任务的描述达成一致。这种严格的定义势必会导致制造服务发现与匹配的失败。因此，制造服务匹配算法需要适应一个较为宽松的“充分相似”的定义，需要有较强的适应性，也就是说，这种匹配算法应该能够依据协同制造链发起企业定义的匹配度进行匹配。如上文制造服务匹配引擎工

作流程所述,协同制造链发起企业在进行制造服务匹配之前应该首先确定其需要的匹配程度,并将其提交给制造服务匹配引擎。

制造服务匹配度在一定程度上反映了制造服务发现与匹配结果的精确度。协同制造链发起企业为制造服务匹配设定一个匹配度后,制造服务匹配引擎将把该匹配度等级之上的制造服务全部搜索出来,如用户设定匹配度为 9,则匹配引擎进行匹配时,匹配度为 10、11、12 的制造服务也将被搜索出来,而且同一匹配度等级内部可能会搜索出多个满足要求的制造服务。为了衡量制造服务制造能力特征与协同制造任务制造能力需求之间的密切程度,对于上述匹配结果,通常需要进行进一步的排序。为此,本文引入了制造服务语义相似度的概念,制造服务语义相似度主要用于描述制造服务和协同制造任务在制造能力层次上的语义相似度,从而为制造服务匹配结果排序提供一个量化的标准。

由第 3 章论述可知,制造服务与协同制造任务均使用网络协同制造本体中定义的一系列概念进行描述。因此,制造服务语义相似度可以通过计算网络协同制造本体中概念间的语义相似度来获得。本文通过建立函数 Semsimilarity(T,S)来计算制造服务语义相似度,具体见下式:

$$\text{Semsimilarity}(T,S)=\frac{\omega_1\text{SemS}(C_T^P,C_S^P)+\omega_2\text{SemS}(C_T^S,C_S^S)+\omega_3\text{SemS}(C_T^M,C_S^M)}{\omega_1+\omega_2+\omega_3}\in[0\cdots1]$$

其中,T、S 分别表示协同制造任务与制造服务;$\text{SemS}(C_T^P,C_S^P)$为协同制造任务与制造服务的零件类别概念语义相似度计算函数;$\text{SemS}(C_T^S,C_S^S)$为协同制造任务与制造服务的形状特征概念语义相似度计算函数;$\text{SemS}(C_T^M,C_S^M)$为协同制造任务与制造服务的材料特征概念语义相似度计算函数:$\omega_1,\omega_2,\omega_3\in[0\cdots1]$分别表示零件类别概念、形状特征概念以及材料特征概念在制造服务匹配过程中的权重。由于在制造服务匹配过程中,加工类型必须一致,故上式中没有包含加工类型的语义相似度计算。

在同一本体中,两个概念 C_i 和 C_j 之间的语义相似度≤1。当两个概念相一致的时候,即具有 Equivalent 关系时,两者之间的语义相似度等于 1;而当两个概念具有 Fail 关系时,两者之间的语义相似度等于 0;对于介于上面两种情况之间的概念,即概念之间具有 Subconcept 或 Relative 关系时,概念之间的语义相似度需要通过计算求出。概念之间的语义相似度描述见下式:

$$\text{Sem}=\begin{cases}1\\ \text{Similarity}(C_i,C_j)\text{Subconcept}(C_i,C_j),\text{Relative}(C_i,C_j)\\ 0\end{cases}$$

目前,一种比较直观的计算概念间语义相似度的方法是将两个概念分别映射到本体后,计算本体图上两个概念节点间的最短路径,但计算图上节点间的最短距离复杂度较高,采用 Dijkstra 算法和 Floyd 算法的复杂度分别为 $O(n^3)$和 $O(n^2)$。本文计算概念之间的相似度主要依据 Tversky 的基本特征相似性模型进行,该模型被认为是迄今为止最有效的计算概念之间相似度的模型。

Tversk 的模型基于如下思想:Tversky 将评估两个概念相似性的特征分为共同特征和不同特征两种。共同特征能够增强两个概念的相似性,而不同特征则会减弱相似性,但是共同特征对相似度的增强影响要大于不同特征减弱相似度的影响。所以在评价相似度的时候,相对于不同特征而言,我们会给予概念的共同特征以更大的信任度。举个例子,比如说

赛车和轿车，它们非常相似，因为它们有很多共同特征，如车轮、引擎、方向盘、排档等。但是它们又有一些区别让它们不相似，比如高度和轮胎的尺寸等。在网络协同制造本体中，概念的特征主要通过属性来体现，因而概念之间特征的比较可以通过对概念之间属性的比较来实现。基于 Tversky 的模型，函数 Similarity(C_i, C_j)的语义相似度计算如下式所示：

$$\text{Similarity}(C_i, C_j) = \sqrt{\frac{|P(C_i) \cap P(C_j)|}{|P(C_i) \cup P(C_j)|} \times \frac{|P(C_i) \cap P(C_j)|}{|P(C_j)|}}$$

其中，函数 $P(x)$表示与概念相关的所有属性，函数$|x|$则返回 x 中属性元素的个数。函数 Similarity(C_i, C_j)的结果由两个部分的几何平均值得到：两个概念的共同属性占两个概念所有属性的比率，两个概念的共同属性占被匹配概念的所有属性的比率。

12.6.2　制造服务资源访问策略

智能协同制造服务网络中个体、智能工厂在异地存放有大量的制造资源、中间成果，例如委托 IME-A 开展设计，委托 IME-B 准备原材料，委托 IME-C 进行加工，委托 IME-D 代理销售，不同制造应用之间服务、资源的访问都将被操作者需要，不同智能工厂间应用或服务的安全机制我们称为智能协同制造服务体系下的服务联合安全与授权，涉及两个方面：身份认证和服务授权。PKI(Public Key Infrastructure)体系是目前单 IME 建立内部不同应用之间统一身份认证的通用手段。基于 X.509 协议的 CA 是 PKI 的核心。CA 中心签发 CA 证书。现行的 PKI 机制一般为双证书机制，即一个实体应具有两个证书，两个密钥对，分别用于加密和签名。CA(Certificate Authority)中心及应用集成是 PKI 体系的实现，一个完整的 PKI 体系包括根 CA、子 CA 中心、密钥管理服务器、证书签发服务器、安全审计服务器、证书目录服务器、注册服务器、OCSP(Online Certificate Status Protocol)服务器、远程注册系统、证书审批服务器。企业利用 CA 系统中的应用 API 和安全服务 API 实现统一身份认证和单点登录。

12.6.3　制造服务资源的优化与智能调度

在智能协同制造服务网络选择中每个智能工厂的生产规模、生产设备、技术专家、普通劳动力、原材料等制造资源时刻处于动态变化中。传统的制造活动会受限于本地资源。

在智能协同服务全域资源可用条件下，制造活动有可能不至于由于某些资源的欠缺导致错失良机。在此类情况下，我们希望找到一种解决方案来从可用制造资源中选择最优方案。

我们将问题抽象如下。

制造资源：智能工厂拥有的智能制造资源简化为设备、专家、工程师、工人、原材料等类别。我们假定一个生产任务通过上述制造资源可以完成。

本地资源：智能工厂自有的智能制造资源。

虚拟资源：通过智能协同制造网络体系可以访问并得到授权的智能制造资源。

假设条件：

(1) 加入到智能协同制造服务的 IME 本身是可信的。

(2) SCIM 中被共享的 IMR 是真实可信的。

(3) 可用虚拟资源已经通过资源分享和同步机制在有限时间内得到确认。

(4) 网络环境是可靠的。

(5) 资源可达。线上所获得的资源可以通过线下的法律商务活动得到确认。

问题：在生产紧张，本地没有可用资源的情况下，如何从虚拟资源中选择最优的合作伙伴。

1. 制造服务资源的优化

如下所示，在紧耦合的智能工厂联盟域 Domain2 中的虚拟资源请求者 IME-A 向本地域服务器发出虚拟资源的请求。制造资源服务器可基于前述 LDAP 服务实现。A 请求资源并进行资源决策的过程如以下流程示意。首先，假定资源仅限于设备、工程师、专家、工人和原材料。各种资源具有更多的参数，例如，设备资源包括其利用成本、生产速度、刀具数量、工艺精度、故障率；工程师可以细分至其专业方向等。需要按照以下步骤来执行。

1) 发送请求

当智能工厂因条件限制，自身所拥有的资源不足以满足当前的生产要求时，可以选择向虚拟组织本地域制造资源服务器发送申请，请求资源调配。此时智能工厂应向服务器提交具体的资源需求标准及限制条件，如资源类型、数量、成本要求，精度要求、时间要求等，同时还应提交对调度资源的偏好信息，以供后期建模寻找最优方案提供参考依据。

2) 数据传递与分析

当资源服务器收到智能工厂提交的请求信息时，基于已经得到收敛的制造资源信息进行计算。智能工厂通过本地资源请求终端提出虚拟资源的跨域边界。

3) 建立模型求解方案

多目标规划研究多于一个的目标函数在给定区域上的最优化。从资源服务器寻找资源匹配方案时需要考虑多方面因素，对此需要建立多目标规划模型。针对现实问题的复杂性，在效用最优化模型(又称线性加权法)与隐枚举法的基础之上，寻找最优方案。将多个目标函数根据申请方提供的参数权重为系数建立效用函数将多目标问题转化到传统的单目标规划问题。运用隐枚举法，找到满足要求的可行域，对于不落入可行域的解筛除，获得多组非劣解，以效用函数作为评判优劣好坏的依据，对可行解进行排序、组合，获得若干种解决方案。

对于每笔订单，我们总希望能够以尽量少的成本得到品质尽量好的产品，所以设目标函数为：

$$F(\boldsymbol{X})=\begin{bmatrix}\min f_1(\boldsymbol{X})\\ \max f_2(\boldsymbol{X})\end{bmatrix}$$

式中，$\boldsymbol{X}$ 为 n 维决策变量向量。

第一个目标函数表示的是所求的解对应的生产成本应尽可能低，第二个目标函数表示的是所求的解对应的产品工艺应尽可能高。

根据各个工厂的生产因素数据定义系数矩阵 $\boldsymbol{A}_{T_p}$：

$$\boldsymbol{A}_{T_p}=\begin{bmatrix}a_{11} & a_{12} & h(1,p) & a_{14}\\ a_{21} & a_{22} & h(2,p) & a_{24}\\ \vdots & \vdots & \vdots & \vdots\\ a_{n1} & a_{n2} & h(n,p) & a_{n4}\end{bmatrix}$$

$\boldsymbol{A}_{T_p}$ 表示为 p 时刻所对应的各工厂的生产参数。

现需要生产产品 Y，x 千件，假设对于工厂 i 完成这笔订单所需时间为

$$t_i(t_i \in Z, i = 1, 2, \cdots)$$

那么满足 t_i 满足条件式

$$\sum_{p=0}^{t_i} a_{i2} h(i, p) \geqslant x$$

在模型中定义 t_i 为最小的时间单位，往下不再可分，作为离散问题的解保证 t_i 为整数，故 t_i 应取满足上述条件式的最小整数。

通过以上条件求得 t_i 并进一步得到新的系数矩阵 $\boldsymbol{A}_1$

$$\boldsymbol{A}_1 = \begin{bmatrix} a_{11} & a_{14} & t_1 \\ a_{21} & a_{24} & t_2 \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n4} & t_n \end{bmatrix}$$

在具体的实际问题中，对于生产订单经常会有一些硬性限制条件，比如该订单需要在多长时间生产完成、生产成本要在多少以下、产品质量要达到一个什么标准，对此，我们引入约束向量

$$b = (b_1, b_2, b_3)$$

其中：b_1 为生产成本约束；b_2 为生产工艺约束；b_3 为生产时间约束。

这样就可以建立如下标准的多目标线性规划模型（LP*）

$$F(\boldsymbol{X}) = \begin{bmatrix} \min f_1(\boldsymbol{X}) \\ \max f_2(\boldsymbol{X}) \end{bmatrix}$$

$$\text{s.t.} \begin{bmatrix} a_{11} & a_{14} & t_1 \\ a_{21} & a_{24} & t_2 \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n4} & t_n \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} < \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

式中：$f_1(\boldsymbol{X})$ 表示生产成本，$f_2(\boldsymbol{X})$ 表示产品工艺，$\boldsymbol{X} = [x_1, x_2, \cdots, x_n]^{\mathrm{T}}$ 为决策变量向量；$\boldsymbol{A}_1 = \begin{bmatrix} a_{11} & a_{14} & t_1 \\ a_{21} & a_{24} & t_2 \\ \vdots & \vdots & \vdots \\ a_{n1} & a_{n4} & t_n \end{bmatrix}$ 为系数矩阵；$\boldsymbol{b} = [b_1, b_2, b_3]^{\mathrm{T}}$ 为约束向量。

传统的多目标线性规划问题，最终只从 n 种决策中选择一种作为问题最优解，现考虑有决策组合的可能性，即可将多种选择组合至一起形成新的可行解。

2. 智能调度

当我们已确定了最终资源组合方案，在最终资源组合方案中，向本地提供空闲资源的工厂称为合作伙伴。假定具体的生产工序已知，此时考虑一种特殊情形：智能协同制造网络体系中合作伙伴所提供的设备资源具有柔性，可负责完成生产流程中的多种工序。

此时，整个生产流程并非会全部在本地工厂内部完成，而是将生产流程根据资源配置

及具体工序划分成若干子流程分配给各个合作伙伴，形成 VPL。为缩短时间，可能需要同一时刻一起进行多个子流程。下面将针对在资源池已经完备且要同时进行多个子流程作业的情况下，如何将资源合理分布到各个合作伙伴进行研究。为方便分析，将问题抽象如下。

假设条件：

（1）从合作伙伴处得到的虚拟资源刚好满足生产的资源需求量，无多余资源。

（2）生产流程是可分的，根据具体生产工序可分为多个部分。

（3）合作伙伴的设备资源均具有柔性，能够负责完成多种子流程。

（4）所讨论的所有子流程为同时刻进行的。例如，造船的分段工作。

（5）任何一个合作伙伴 IME 在完成任何一个可以完成的子流程时，生产总成本与使用的柔性资源数量成线性比例关系。

（6）生产工序一旦开始进行加工，中途即不再有任何意外情况使其中断。

问题：在生产资源已备齐的情况下，如何合理分配虚拟柔性资源从而使得成本最低？

根据具体的生产任务情况，具体产品所对应的具体生产工序以及申请方对生产任务的时间约束，计算出生产任务中每个子流程所需要的资源数量，设备资源虽为柔性，但对于不同生产子流程，其作业成本与速度不尽相同，考虑到地域问题，运输成本也存在差异。综合考虑合作伙伴的生产能力、人力资源、公司规模、地理位置等情况，物料运输成本和产品运输成本，计算出每个合作伙伴单位资源的生产成本，假定单位生产成本与使用资源数量成正比例关系，针对分配问题，利用产销平衡问题（平衡运输问题）的解题思想，建立模型。

遵循总成本最低的原则寻找分配方案，对此我们建立分配问题数学模型

$$\min Z=\sum_{i=1}^{n}\sum_{j=1}^{m}c_{ij}x_{ij}$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}x_{ij}=b_j, & j=1,2,\cdots\\ \sum_{j=1}^{n}x_{ij}=a, & i=1,2,\cdots\\ x_{ij}\geqslant 0\end{cases}$$

单纯形法的基本思想是：先找出一个基本可行解，对它进行鉴别，看是否是最优解；若不是，则按照一定法则转换到另一改进的基本可行解，再鉴别；若仍不是，则再转换，按此重复进行。因基本可行解的个数有限，故经有限次转换必能得出问题的最优解。如果问题无最优解也可用此法判别。

具体计算步骤归纳为以下几点：

（1）确定初始基可行解，将基可行解填入资源分配表。

（2）用闭回路法求出各个非基变量的检验数，并判别该基可行解是否已达到最优解。若已经是最优解，则停止计算，若不是，执行下一步。

（3）用闭回路法进行调整。确定入基变量、出基变量，寻找新的基可行解。

（4）重复步骤（2）和步骤（3）一直到求出最优解。

12.6.4　智能化协同制造研究与自学习机制

在智能协同制造生命周期过程中，市场研发、制造服务各环节的自我更新是智能工厂的重要活动组成。智能工厂必须从前沿技术、专家智慧和竞争对手的创新中获得灵感。智能工厂作为一个复杂的智能体，其自我演化的智慧应该来自于以下 4 个方面。

（1）来自企业内部业务系统的结构化或者非结构化数据。结构化数据可以通过可视化的报表工具来进行商业智能挖掘。

（2）内部海量的制造业务文档的社会化关系所隐含的智慧。

（3）企业外部制造联盟的关联文档及社会化关系所隐含的智慧。

（4）来自整个智能协同制造服务的关联技术文档及社会化关系所隐含的智慧。

但是，智能工厂内的制造智慧散落在各个制造应用中，且存在如下一系列问题。

问题一，通常情况下，企业制造应用包括面向特定专业的资讯、技术文档、推送消息，这些内容往往包括在制造应用的“帮助”“消息”“看板”等不同的功能模块中。实际情况是制造应用系统在企业内已经形成了矩阵的应用模式。每个应用都在向使用者发送具有一定重叠和交叉的知识，但是并没有形成关联关系。

问题二，智能工厂内的制造服务应用经常基于不同操作系统平台，不同数据格式，如 MySQL，Microsoft SQL，Oracle，DB2 等不同数据库格式。互联网思想影响智能工厂内的应用，制造服务应用的 UI 也与往常的 MIS 大不一样，更多的页面交互性往往都带有验证码，社区论坛的逐层消息也和传统的二维数据库关系有所差异。CAD、CAM、CAPP 等专有制造文档，由于在线培训的需要，MP3、MP4、JPG、PNG、BMP 等多媒体文档也成为制造知识的一部分，但经常游离在外。协同制造生命周期所包含的产品体验数据来自互联网，通常以微博、微信、QQ 的方式反馈，它们也是制造知识来源的一部分。

问题三，本地文件、异构的分布式文件系统使得制造文档的存储分布在不同地方产生了文件统一搜索问题。

问题四，智能工厂内以及智能工厂间的知识没有得到社会化共享。举例来说，智能工厂内的设计部门员工 A 需要了解“元器件 X 的耐腐蚀性材料的新一代技术”，其实企业集团的制造部门的员工 B 由于个人感兴趣已经收集了大量材料，由于双方之间没有形成协同，员工 A 需要重新探索一次。同样的场景也会发生在智能工厂间而产生信息距离。以上信息源大数据，通过信息采集和分析处理进行知识发现，如图 12-18 所示。

为解决上述问题，我们将智能化协同制造服务内的所有终端都看作可进化的种群，终端本身是弹性的，既可以小到员工、制造服务应用，也包括智能工厂本身或者一个紧耦合的智能工厂联盟，构造智能协同制造服务的制造智慧遗传算法。基于遗传算法，协同制造服务体系本身、单智能工厂、单智能工厂内的制造应用、智能工厂内员工均可以实现自学习。遗传算法允许知识自我更新的人工干预和自动配置的方式。基于遗传算法的知识学习可以作为独立的制造应用，也可以通过智能工厂内部数据总线嵌入到制造业务应用中。

其中，要素抽取从非结构化文档中抽取出符合条件的内容，形成字段。行业关键字规则预先建立行业的常用关键字及其逻辑推理关系，用于之后的推理机制。主题知识则预先初始化常用主题知识库。

格式识别基于文件头信息识别，将无意义的内容清洗后，保留核心数据，创建统一的

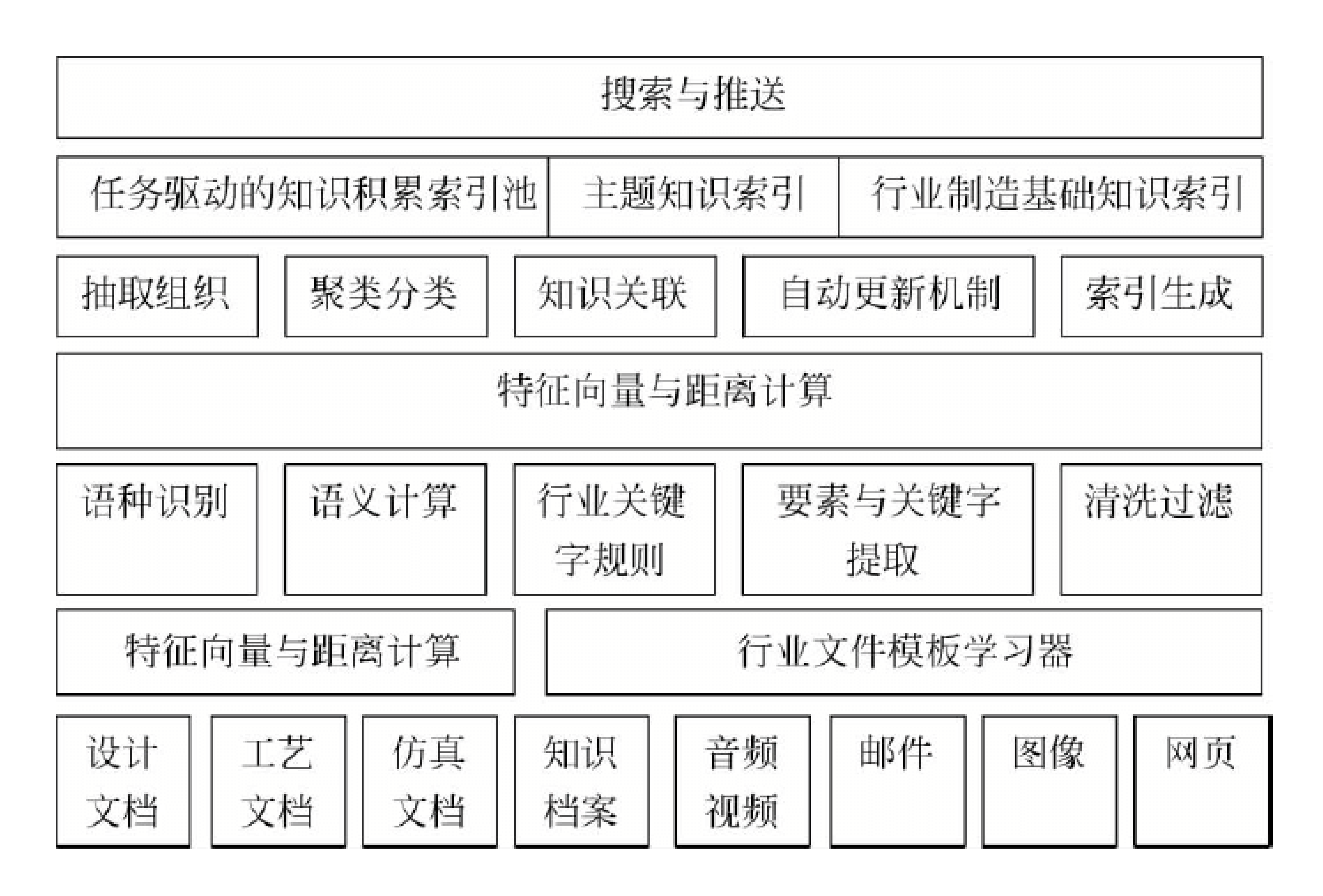


图 12-18　文件内容采集与处理流程的层次结构

XML 格式的索引。结合基于距离的语义相似度、基于信息内容的语义相似度、基于属性的语义相似度、混合式语义相似度算法构造混合语义分析算法，可以相对较为准确地理解制造行业技术文档及相关自然语言的语义。混合算法的形式化表示如图 12-19 所示。

(Sim(C1,C2)*p1+Sim(M1,M2)*p2+Sim(L1,L2)*p3-Dis(M1-M2)*p4-Dis(C1-C2)*p5)/P6*Y1+ (Sim(S1,S2)*p7-Dis(S1,S2)*p8) *Y1+ (Sim(D1,D2)*p9-Dis(D1,D2)*p10) *Y1+ (Sim(V1,V2)*p11-Dis(V1,V2)*p12) *Y1

图 12-19　语义分析算法的形式化公式

在上述公式中，C1、C2 代表距离计算数值，M1、M2 代表高频词统计，L1、L2 代表低频词统计，S1、S2 代表句子统计，D1、D2 代表段落统计，V1、V2 代表谓词统计，P1、P2、P3、P4、P5、P6、P7、P8、P9、P10、P11、P12 均代表可调参数，在域内初始运行时刻经过统计计算出来，可根据结果人工调整。Y1 代表语言学可调参数，与语种有关，不同的语种具有不同的参数。Sim 是相似性计算，Dis 是差异性计算。

用户输入想要查找的样本或者特征，搜索进程学习目标的特征，生成特征规则，然后利用特征规则在文件中对内容进行比对，最后发现未知内容。通过复合选择项参数：精确匹配、相似、前精确后相似、前相似后精确、特殊部位相似来配置比对条件。比如输入“F22/A”，选择“前精确后相似”，那么根据上述机制可生成特征公式：F[-]＊22[/-]＋?[\w]＋?，这时候可以发现 F22 的各种型号：F22/S、F22/A、F-22/A、F-22/B 等。如果用户选择“相似”，那么可生成另一特征公式：[\w]＋?[-]＊[\d]＋?[/-]＋?[\w]＋?，这时候不但能发现 F22/A，而且还能自动发现 F35/A、F35-B、F-35/A、Su35-B 等符合特征的近似内容。构造基于特征学习的未知内容发现机制包含特征发现、特征学习、模式匹配构成。特征发现是对每一个词的前词、后词、相似性、变化性进行统计，继而发现一些特征经常出现的概率，从而发现特征并计算特征可能性的值。发现特征后，记录下特征的变化规律，并自动生成模式公式，每一种特征均有一组模式公式相对应。模式匹配可采用 KMP 匹配算法和 BM 匹配算法。推理检索机制的目的是解决文章语义包含某关键字方面的意思，但明文不包含关键字的情况。推理是用本体推理机完成的，本体提供共享词表，即特定制造领域之中那些存在着的对象类型或概念及其属性和相互关系。一旦通过语义提取并进行知识关联后，除了常规关键字检索，还可通过预先设定的逻辑规则。

概括而言，本地域中指定的开放数据源、在本地的应用数据将在智能协同制造企业域内挖掘到并交换给同域内的其他代理，基于 JXTA 的 P2P 网络上的代理服务器将交换彼此的智慧从而实现智能协同制造(ICM)体系内的制造智慧的遗传进化。

12.7　工业大数据展望

随着工业互联网建设和应用不断深入，数据的价值与作用将越来越凸显，数据分析将向工业各环节渗透，预测、决策、控制等更智能的应用成为发展方向，最终构成从数据采集到设备、生产现场及企业运营管理优化的闭环。工业数据未来将呈现出以下几个发展方向：一是跨层次跨环节的数据整合。当前工业数据水平来看分散在研发设计、生产管理、企业经营等各个环节，垂直来看分散在生产现场、企业管理(MES、ERP)等不同层次，下一步数据在垂直和水平两个方向都需要整合，为全局视图分析奠定数据基础。其中，语义技术将发挥重要作用，利用语义可以对工业互联网数据的含义进行标注，使数据在异构系统之间能够被正确理解和处理。二是数据在边缘的智能处理。在靠近数据源头的网络边缘节点上，通过融合计算、存储与控制等功能，实现数据的边缘处理、分析与过滤，以满足工业生产现场实时连接、实时控制、实时分析、安全隐私等需求，并可以与云平台实现互补。三是基于云平台数据集成管理。将数据汇聚起来，上传到云计算平台进行分析处理，是未来的主流方向，基于成熟的、经验证的技术以及大数据平台来支撑工业数据的数据建模、数据抽取 ETL、查询与计算，与传统实时数据库、关系数据库和 MPP 数据混搭应用，是云化的工业大数据平台构建的主流方向。四是深度数据分析挖掘。知识驱动的分析方法，建立在工业系统的物理化学原理、工艺及管理经验等知识之上。数据驱动的分析方法，完全在数据空间中通过算法寻找规律和知识。未来的发展趋势是更多地将基于知识的方法与数据驱动方法融合，满足工业数据分析对高置信度的要求。五是数据可视化。建立机器、生产流程、全生产周期等拟真数字化模型，并进行可视化呈现，使生产管理者、系统开发者和用户能够更加直观全面地了解相关信息，支撑设计、生产、产品流通与交易、产品服务等环节的决策水平。

同时，网络化制造技术是一个充分融合制造技术与信息技术的前沿研究方向，因此充分发挥信息技术的优势是其不断发展和取得突破的方向所在。本章在写作时注意将当前信息技术领域(语义技术、Web 服务、网格计算等)的新思想、新成果与网络化制造技术的研究进行了紧密结合，针对智能工厂异地协同制造过程，提出了智能协同制造服务概念，目的是希望寻找一种通过虚拟生产线使智能工厂间合作制造的网络化制造实现方式，为实现智能化协同制造服务的大规模工业化应用做一些积极有益的探索。



大数据工程保障体系建设

大数据工程建设是一项巨大的系统工程，它涉及社会或各行各业信息化数据的综合集成和知识发现。其中，政府、金融、环保、医疗、工业制造等又属于热点行业而且数据交互强、数据关联度高、涉及社会多方面，为了切实有效地推进大数据工程建设，必须进一步建立健全的相关保障机制。

13.1　法律体系建设

大数据环境下，企业信用呈现体态虚拟化与数字化、影响因素广泛化与纵深化的新特点，而企业信用监管的法律、制度不健全，相关保障措施欠缺。应建立以企业信用基本法为基础的企业信用监管法律体系，并以中央为主导、地方为特色完善企业信用分类监管制度，提升大数据技术处理能力与企业信用监管水平，注重企业信用法律监管中权益保护的均衡性，做好企业信用法律监管的保障工作。

计算机技术的发展和云计算技术的兴起使得大数据在社会经济生活中的应用不断加强与深化。大数据是一个相对比较抽象的概念，单是从字面来看就表示了数据之多之大，但其最主要的内涵在于数据的全面性和不可穷尽性。截至目前，学界尚未对大数据形成统一的概念。一般认为，大数据是指数量巨大、类型众多、结构复杂、有一定联系的各种数据所构成的数据集合。大数据的主要功能在于可以不断提升数据的使用价值，实现数据的快速流转和多样化的数据处理模式。大数据为企业的经营决策提供了更为全面详尽的数据支持，为企业的信用信誉建设搭建了新的平台和快速构建通道。大数据技术的不断发展势必会对企业信用监管体系产生极大的挑战与冲击，同时也会为其发展革新带来新的机遇，如何更好地迎接挑战，把握机遇就显得尤为重要。

1. 构建全方位、立体化的法律监管

(1) 建立以企业信用基本法为基础的企业信用监管法律体系大数据环境下，企业信用监管呈现出新的特点，需要更为细致完备的法律去对监管的各个环节进行规制，从而实现新环境下监管工作有法可依的状态。完善的法律监管模式应在包括消费信用、工商信用以及信贷等有关信用交易体系内形成全方位的、严密的监管法律。信用交易可以极大地便捷市场交易行为及扩大市场交易规模，有效地适应全球化贸易的需要。良性高效运行的信用交易必须形成于国家信用管理制度之上，而要形成健全的国家信用监管体系就必须健全信用监管的法律，完善立法。在很大程度上，企业主个人的信用行为会影响企业信用，所以，立法

应该将企业主的个人行为纳入企业信用监管体系内，对其进行并列监管并以个人信用行为为限对企业信用违法行为承担连带责任。大数据环境下，国家最高权力机关更应加快制定规范企业主体信用行为，调整各个信用主体间权利义务关系的信用基本法。通过立法对相关主体的权利义务予以明确，为信用数据的收集、处理以及各主体信用行为的奖惩评判提供法律依据，在信用数据的来源、存储、使用的过程中实现全方位、立体化的监管。同时，应当在结合本国实际的前提下积极借鉴欧美等信用法制发达国家的有关立法经验，制定出可行性强、有效性高的本国信用监管法律。企业信用的相关者众多且各相关者所提供的有关该企业信用的数据是对企业进行信用监管的重要数据依据。信用表现为对民事主体经济信赖的社会评价，信用的客观表现是一种评价，这种评价是社会公众的评价，而不是当事人的自我经济评价；这种评价是对特定主体经济信赖的客观评价，它可能是但不一定是肯定性的社会评价。在企业信用监管立法过程中要坚决贯彻诚实信用原则，诚实信用被奉为民法的基本原则，有“君临法域”的效力。我国《民法通则》、《合同法》中都明确规定了诚实信用原则是市场经济主体进行交易订立合同的基本原则，这就可以明确该原则同时也应成为建立企业信用基本法的基本原则。

（2）以中央为主导、地方为特色完善企业信用分类监管制度在完善企业信用监管法律的基础上，要在日常监管工作中实现对主体信用监管的法制化、常态化，就必须在中央政府的主导下形成全国性的、部门性的及地方性的可执行性强的企业信用监管制度，以彰显企业信用法律监管的实效。如将企业投招标等生产经营行为与企业信用记录结合，对信用数据记录不良的企业市场行为进行必要的限制；将企业失信数据进行累加并明确对失信企业的整改措施等。各地在中央的统一部署下应结合本地域特点完善地方信用监管机制，可根据本地的经济发展水平制定出地方性的企业信用激励机制，对信用良好、诚信度高的企业在制度允许的范围内予以税收优惠、财政补贴等倾斜；同时，应积极建立企业信用不良记录黑名单制度，对信用不良企业予以惩处并曝光，在全社会范围内营造守信获益、失信受损的氛围，以进一步激励企业乃至个人珍视信用，诚实守信。对企业信用分类监管制度进行完善，首先要充分利用大数据的优势，完善企业主体信用数据信息。当前，金融机构对个人信用信息的构建是比较完善的，在对企业信用信息的完善过程中可利用金融机构所具有的个人信用信息，对企业主、企业负责人、法定代表人、股东等与企业信用密切相关的个人信息进行收集融合。其次，在信用监管的过程中应对监管等级进行分类细化。对企业信用等级可采取平级制方法，分别设立 A、B、C 不同的信用等级，对企业信用进行量化管理，激发企业自主地进行诚信建设。

2. 做好大数据环境下企业信用法律监管的保障工作

提升大数据技术处理能力与企业信用监管水平大数据环境下，要实现海量数据的有效整合，挖掘数据信息提升信息价值，就必须进行多种技术的协同。数据挖掘与收集、处理及分析是大数据下企业信用数据处理的主要过程，对数据进行挖掘、存储、使用时必然会涉及引擎搜索技术、云计算处理技术以及数据库技术等一系列的高新技术。所以，在大数据环境下要对企业信用进行高效监管，必须增强学习意识和技术观念，提高自身技能，才能对不法企业运用大数据技术扰乱信用监管秩序的行为进行有效监管，实现有的放矢，堵住不法企业钻技术漏洞的空子。同时，监管过程中还应提升根据现有数据对企业未来信用行为的预测能力，实现对企业信用动态的准确把握，防患于未然，将不法行为扼杀于萌芽状态，引导企业向着健康的方向发展。

13.2 标准体系建设

目前,大数据技术相关标准的研制还处于起步阶段,本部分对 ISO/IEC、ITt 等国际标准化组织、NIST、国内全国信息技术标准会技术委员会已经开展的标准化工作进行梳理,依据大数据技术体系,从基础、技术、产品、应用等不同角度进行分析,形成了大数据标准体系框架。对我国现有标准、在研标准和将提出的标准计划进行分析,形成大数据标准体系。对于目前急需研制的标准进行了较为详细的分析,这部分将成为后续标准化工作的重点。

大数据标准体系是为实现大数据领域的标准化而形成的体系,凡是与此目的有关的大数据领域标准之间都存在相互依存、相互衔接、相互补先、相互制约的内在联系,最终形成科学的有机整体。因此,要求建立的标准体系具有先进性,在应用系统科学理论和方法的基础上,运用标准化的工作原理,着眼于寻找整套的标准内容,基于这些内容,在标准体系的内在联系上进行统一、简化、协调和优化等处理,力求体现出系统内标准的最佳秩序,防止在标准之间存在不配套、不协调、互相矛盾及组成不合理问题。

1. 大数据标准体系框架

结合国内外大数据标准化情况、国内大数据技术发展现状、大数据参考架构及标准化需求,根据数据全周期处理,数据自身标准化特点,当前各领域推动大数据应用的初步实践,以及未来大数据发展的趋势,我国提出了大数据标准体系框架。

大数据标准体系由 7 个类别的标准组成,分别为:基础标准、数据标准、技术标准、平台和工具标准、管理标准、安全和隐私标准、行业应用标准。

(1) 基础标准。为整个标准体系提供包括总则、术语、参考模型等基础性标准。

(2) 数据标准。该类标准主要针对底层数据相关要素进行规范。包括数据资源和数据交换共享两部分,其中,数据资源包括元数据、数据元素、数据字典和数据目录等,数据交换共享包括数据交易和数据开放共享相关标准。

(3) 技术标准。该类标准主要针对大数据相关技术进行规范。包括大数据集描述、大数据处理生命周期技术和操作技术三类标准。其中,大数据集描述主要针对描述模型、分类方法、质量模型和数据溯源等方面进行规范。大数据处理生命周期技术主要针对数据的收集、预处理、分析、可视化、访问等进行规范。

(4) 平台和工具标准。该类标准主要针对大数据相关平台和工具进行规范,包括系统级产品和工具级产品两类,其中工具及产品包括平台基础设施、预处理类产品、存储类产品、分布式计算工具、数据库产品、应用分析智能工具、平台管理工具类产品的技术、功能、接口等进行规范。相应的测试规范针对相关产品和平台给出测试方法和要求。

(5) 管理标准。管理标准作为数据标准的支撑体系,贯穿于数据整个生命周期的各个阶段。该部分主要是对数据管理、运维管理和评估三个层次进行规范。

(6) 安全和隐私标准。数据安全和隐私保护同样作为数据标准的支撑体系,贯穿于数据整个生命周期的各个阶段。抛开传统的网络安全和系统安全,大数据时代下的数据安全标准主要包括方法指导、检测评估和要求三类标准。

(7) 行业应用标准。行业应用类标准主要针对大数据为各个行业所能提供的服务角度

出发制定的规范。该类标准指的是各领域根据其领域特性产生的专用数据标准，包括工业、电子商务、健康等领域。

2. 大数据相关标准明细

根据大数据标准体系框架，整理出发布、报批、立项、申报、在研以及计划的大数据相关国家标准 99 项，大数据标准明细如表 13-1 所示。

表 13-1　大数据标准明细

序号	一级分类	二级分类	国家标准编号	标准名称	采用标准号及采用程度	状态
1	基础	总则		信息技术 大数据标准化指南		计划
2		术语	20141191-T-469	信息技术 大数据 术语		在研
3		参考架构	20141190-T-469	信息技术 大数据 技术参考模型		在研
4				信息技术 大数据 参考架构 第 1 部分 框架和应用指南		计划
5				信息技术 大数据 参考架构 第 2 部分 用例和需求		计划
6				信息技术 大数据 参考架构 第 5 部分 标准路线图		计划
7				信息技术 大数据 基于参考架构下的接口框架		计划
8	数据	数据资源	GB/T 28821—1012	信息技术 数据元素值格式记法	ISO/IEC 14957：1996，IDT	发布
9			20101507-T-469	信息技术 数据元素值表示——格式记法	修订 GB/T 18142—2000：ISO/IEC FDIS 14957：2009	报批
10			GB/T 18391.1—2009	信息技术元数据注册系统(MDR)第 1 部分：框架	ISO/IEC 11179—1：2004，IDT	发布
11			GB/T 18391.2—2009	信息技术元数据注册系统(MDR)第 2 部分：分类	ISO/IEC 11179—2：2005，IDT	发布
12			GB/T 18391.3—2009	信息技术元数据注册系统(MDR)第 3 部分：注册系统元模型与基本属性	ISO/IEC 11179—3：2003，IDT	发布
13			GB/T 18391.4—2009	信息技术元数据注册系统(MDR)第 4 部分：数据定义的形成	ISO/IEC 11179—4：2004，IDT	发布
14			GB/T 18391.5—2009	信息技术元数据注册系统(MDR)第 5 部分：命名和标识原则	ISO/IEC 11179—5：2005，IDT	发布
15			GB/T 18391.6—2009	信息技术元数据注册系统(MDR)第 6 部分：注册	ISO/IEC 11179—6：2005，IDT	发布
16			GB/Z 21025—2007	XML 使用指南		发布
17			GB/T 23824.1—2009	信息技术实现元数据注册系统内容一致性的规程第 1 部分：数据元	ISO/IEC TR 20943—1：2003，IDT	发布

续表

序号	一级分类	二级分类	国家标准编号	标准名称	采用标准号及采用程度	状态
18	数据	数据资料	GB/T 23824.3—2009	信息技术实现元数据注册 系统内容一致性的规程 第3部分：值域	ISO/IEC TR 20943—3：2004，IDT	发布
19			GB/T 32392.1—2015	信息技术 互操作性元模型框架(MFI)第1部分：参考模型		发布
20			GB/T 32392.2—2015	信息技术 互操作性元模型框架(MFI)第2部分：核心模型		发布
21			GB/T 32392.3—2015	信息技术 互操作性元模型框架(MFI)第3部分：本体注册元模型		发布
22			GB/T 32392.4—2015	信息技术 互操作性元模型框架(MFI)第4部分：模型映射元模型		发布
23			20132340-T-469	信息技术 互操作性元模型框架(MFI)第5部分：过程模型注册元模型		在研
24			20132341-T-469	信息技术 互操作性元模型框架(MFI)第7部分：服务模型注册元模型		在研
25			20132342-T-469	信息技术 互操作性元模型框架(MFI)第8部分：角色与目标模型注册元模型		在研
26			20132343-T-469	信息技术 互操作性元模型框架(MFI)第9部分：按需模型选择		在研
27			GB/T 30881—2014	信息技术 元数据注册系统(MDR)模块	ISO/IEC 19773:2011	发布
28			GB/T 30880—2014	信息技术 通用逻辑(CL)：基于逻辑的语言族框架	ISO/IEC 24707:2007	发布
29			2010-3325T-SJ	信息技术 元数据 属性		在研
30		交换共享		信息技术 大数据 开放数据集基本要求		计划
31				信息技术 大数据 开放数据集标识管理		计划
32				信息技术 大数据 开放共享第1部分：总则		计划
33				信息技术 大数据 开放共享第2部分：政府数据开放共享基本技术要求		计划
34				信息技术 大数据 开放共享第3部分：开放程度评价		计划

续表

序号	一级分类	二级分类	国家标准编号	标准名称	采用标准号及采用程度	状态
35	数据	交换共享		信息技术 大数据 开放共享第4部分：政府资源目录体系		计划
36			20141201-T-469	信息技术 数据交易服务平台通用功能要求		在研
37			20141200-T-469	信息技术 数据交易平台 交易数据描述		在研
38				信息技术 数据交易 通用概念描述		计划
39				信息技术 数据交易 交易流程描述		计划
40				信息技术 数据交易 数据管理规范		计划
41				信息技术 数据交易 技术规范		计划
42				信息技术 数据交易 风险评估		计划
43				信息技术 数据交易 交易质量评估		计划
44				信息技术 数据交易 数据价值评估指引		计划
45		大数据描述	20141172-T-469	多媒体数据语义描述要求		在研
46				信息技术 大数据 分类指南		计划
47			20141203-T-469	信息技术 数据质量评价指标		在研
48				信息技术 数据质量检测		计划
49			20141194-T-469	信息技术 科学数据引用		在研
50			20141202-T-469	信息技术 数据溯源描述模型		在研
51		处理生命周期技术	20141204-T-469	信息技术 通用数据导入接口规范		在研
52				信息技术 通用数据导入接口测试规范		计划
53				信息技术 大数据分析总体技术要求		计划
54				信息技术 大数据可视化工具通用要求		计划
55			GB/T-12991—2008	信息技术 数据库语言 SQL 第1部：框架	ISO/IEC 9075—1:2003，IDT	发布
56		互操作技术		信息技术大数据互操作技术指南		计划

续表

序号	一级分类	二级分类	国家标准编号	标准名称	采用标准号及采用程度	状态
57		系统级产品		信息技术 大数据 存储与处理系统基本功能要求		计划
58				信息技术 大数据 存储与处理系统功能测试规范		计划
59				信息技术 大数据 分析系统基本功能要求		计划
60				信息技术 大数据 分析系统功能测试规范		计划
61				信息技术 大数据 系统通用规范		计划
62		工具级产品		信息技术 大数据 面向应用的基础计算平台基本性能要求		计划
63			GB/T 28821—1012	关系数据管理系统技术要求		发布
64			GB/T 30994—2014	关系数据库管理系统检测规范		发布
65			GB/T 32633—2016	分布式关系数据库服务接口规范		发布
66			20121409-T-469	非结构化数据表示规范		报批
67			20121410-T-469	非结构化数据访问接口规范		报批
68			GB/T 32633—2016	非结构化数据管理系统技术要求		发布
69			20141183-T-469	实时数据库通用接口规范		在研
70				非结构化数据查询语言		计划
71				智能硬件通用大数据接口规范		计划
72	管理	数据管理		信息技术 大数据 资产管理指南		计划
73		运维管理		信息技术 大数据 系统运维和管理功能要求		计划
74		评估		信息技术 大数据 解决方案基本评估规范		计划
75			20141184-T-469	数据能力成熟度评价模型		在研

续表

序号	一级分类	二级分类	国家标准编号	标 准 名 称	采用标准号及采用程度	状态
76	大数据安全和隐私	要求	GB/T 20009—2005	信息安全技术数据库管理系统安全评估准则		发布
77			GB/T 20273—2006	信息安全技术数据库管理系统安全技术要求		发布
78			GB/T 22080—2008	信息技术安全技术信息安全管理体系要求	ISO/IEC 27001:2005，IDT	发布
79			GB/T 22081—2008	信息技术安全技术信息安全管理实用规则	ISO/IEC 27002:2005，IDT	发布
80			GB/T 31496—2015，IDT	信息技术安全技术信息安全管理体系实施指南	ISO/IEC 27003:2010，IDT	发布
81				信息安全技术 大数据参考架构 第 4 部分 安全和隐私		计划
82				信息安全技术 大数据安全分级指南		计划
83				信息安全技术 大数据安全参考架构		计划
84				信息安全技术 数据脱敏指南		计划
85				信息安全技术 大数据平台安全技术要求		计划
86				信息安全技术 大数据跨集群安全技术框架		计划
87			20130323-T-469	信息安全技术个人信息保护管理要求		在研
88			20130338-T-469	信息安全技术移动智能终端个人信息保护技术要求		在研
89		检查评估		信息安全技术 隐私保护评估方法		计划
90		方法指导		信息安全技术 大数据中的隐私保护框架		计划
91				信息安全技术 个人信息保护指南		立项
92			GB/Z 28828—2012	信息安全技术公共及商用服务信息系统个人信息保护指南		发布
93	行业应用	工业大数据		信息技术 工业大数据 术语		计划
94				信息技术 工业大数据 参考架构		计划
95				信息技术 工业大数据 产品核心元数据规范		计划
96				信息技术 工业大数据 工业订单元数据规范		计划
97		电子商务大数据		信息技术 电子商务大数据 采集规范		计划
98				信息技术 电子商务大数据 仓库模型规范		计划
99				信息技术 电子商务大数据 应用指标体系		计划

通过对现有标准进行梳理可以发现：

(1) 在数据资源方面，我国已经有一些相关标准，同样适用大数据的应用。

(2) 在交换共享方面，由于近年来智慧城市的快速推进，政府大数据进行了广泛而深入的开放共享融合，但数据开放共享方面标准欠缺较多。虽然在研 2 项交易类标准，但是在交易流程和交易数据管理等方面的标准不足。

(3) 在技术标准方面，在数据访问方面，底层数据库标准和数据导入方面已经有相关标准，但数据分析、可视化等缺乏。数据质量是大数据应用和发展的基础，都处于在研阶段。大数据安全方面，虽有基础安全类标准，但缺乏针对大数据的安全框架、隐私、访问控制类标准。

(4) 在大数据平台和工具方面，也主要是在数据库、非结构化数据管理产品类方面已经在研或发布，但在大数据系统级相关产品的标准方面欠缺较多。

总之，针对大数据标准，我国在数据管理、信息安全等方面，已经发布和在研一些标准，具有大数据环境的基础支撑作用，但在整体上缺乏统一规划，相关标准缺失较多。尤其在数据开放共享、数据交易、数据安全、系统级产品等方面，需要尽早补充完善。

13.3 建立标准化大数据治理体系

大数据治理需要建立成熟度模型用于成熟度评估。模型成熟指标需要从以下 11 个方面进行考虑。

(1) 业务成果。代表信息治理计划的目标和目的。

(2) 组织结构和认识。指业务部门和 IT 部门间的相互责任，以及对治理不同管理层次中数据的信托责任的认识。

(3) 管理人员。旨在保证数据监护，实现资产增值、风险消减和组织控制的质量控制准则。

(4) 数据风险管理。据以识别、保留、量化、规避、接受、消减和转嫁风险的方法论。

(5) 政策。期望得到落实的组织行为的书面表达。

(6) 数据质量管理。指测量、提高和保证产品数据、测试数据和归档数据的质量和集成性的方法。

(7) 信息生命周期管理。有关信息采集、使用、保留和删除的系统化的、基于策略的方法。

(8) 信息安全与隐私。组织用于消减风险和保护数据资产的策略、实践和控制手段。

(9) 数据架构。结构化和非结构化数据系统及应用的架构式设计，用户实现数据的可用性，并将数据分配给合适的用户。

(10) 分类和元数据。指用于创建常见的语义定义、IT 术语、数据模型和数据库的方法和工具。

(11) 审计信息日志和报告。指监测和测量数据价值、风险和信息治理有效性的组织流程。

图 13-1 总结了 IBM 信息治理委员会成熟度模型中评估信息治理成熟度的 11 个指标。

可以将上述 11 个指标归纳为以下 4 类。

(1) 目标。指信息治理计划的预期结果。目标倾向于关注降低风险与提升价值，这反

过来又受降低成本和提高收入的驱动。

(2) 支持要素。包括组织结构和认识、管理人员、数据风险管理及政策。

(3) 核心准则。包括数据质量管理、信息生命周期管理，以及信息安全和隐私。

(4) 支持准则。包括数据架构、分类和元数据，以及升级信息日志和报告。



图 13-1　IBM 信息治理委员会的成熟度模型

13.4　加强大数据行业应用研究

大数据行业应用是大数据发展的原动力，基于大数据各个领域、各个层次都提出了丰富多样的应用需求。加强对大数据行业应用需求分析、开展对相关领域需求的研究，展开有针对性应用问题的个性化研究，也要进行业务需求问题的共性研究，行业领域业务需求组织层次、系统框架、逻辑关系、组织关联等行业需求特点研究。梳理整个大数据生态系统脉络体系，把握行业领域重要发展方向，研究行业标准和制度规范。

13.5　加强元数据的研究和应用

在大数据时代，借助于元数据了解数据元素含义和上下文的需求越来越强烈，缺乏统一的数据描述。加强元数据标准或元数据模型的研究和应用，健全完善元数据标准规范及元数据模型。充分结合政府各部分现有数据资源建设情况，针对当前政务大数据资源、工业大数据资源、电子商务大数据资源等重点领域，研制元数据标准或统一的元数据标准模型框架，建立元数据资源库，使得大数据向着标准化、条理化、脉络化方向发展。

13.6　加强大数据核心技术研究

近年来，大数据产业已经成为影响全球未来社会经济发展的战略性新兴领域，应用需求强烈，需要不断完善大数据核心技术体系，突破或改变现有的大数据采集、分析、存储管理等关键技术，加强信息组织和数据仓库研究，形成自主可控的大数据核心技术架构，为大数据

获取、管理和分析提供技术保障。

13.7 促进大数据交易市场的规范化发展

随着大数据技术的成熟和发展，大数据交易市场的建立，大数据在商业上的应用越来越广泛，完善相应的标准及管理制度，规范大数据交易市场，推动行业自律，打造完善、健康、有序的交易产业链条，从交易平台、交易主体、交易对象等多个方面规范交易市场行为，对交易市场内的在线数据交易、离线数据交易、托管数据交易等数据交易模式进行规范。有效解决数据交易中各方的困惑，理顺市场渠道，规范数据交易行为。

13.8 推动大数据标准化进程

大数据标准化工作是支撑大数据产业发展和应用的重要基础。

1. 借助产学研一体化平台，建立标准化体系

做好促进产学研结合的基础工作，包括构建产学研信息沟通平台，举办会议、论坛、项目调研沟通交流等活动，促进产学研各方的信息沟通与交流。要借助产学研一体化信息服务平台，整合高等院校、科研院所、高科技企业等创新资源，在科研机构与企业单位之间搭建一个产学研合作的桥梁。充分发挥产学研用各方力量，调动一切积极因素，加强大数据标准化顶层设计，从国家层面制定大数据标准规范，建立统一的数据标准和技术规范。

2. 推动大数据技术研发、成果转化和知识产权保护

强化大数据标准化意识，完善大数据标准应用环境，推广标准的试点示范，结合重点地区，行业标准化示范区工作，发挥各地方、各领域在大数据标准化工作中的资源优势，不断推动大数据产业标准化和可持续发展。支持企业加强大数据研发投入，采取多种措施提升自主创新能力；支持企业和产业联盟参与承担科技大数据专项、各类科技计划项目等。提高创新主体创造、运用、管理和保护知识产权的能力。政府通过补贴、奖励等措施，支持创新创业主体获得专利权、商标注册和版权登记；支持创新创业主体参与创制标准、成立标准联盟，推动大数据技术标准的产业化应用。

3. 建立人才引进保障机制、促进人才良性循环

完善大数据建设人才、智力和项目相结合的柔性引进机制，畅通人才引进绿色通道。充分发挥物质和荣誉的双重激励作用，创建培养人才、吸引人才、用好人才、留住人才的良好环境。大力培养、引进和高水平使用一批复合型高层次大数据人才、信息专业技术人才、高技能人才、物联网科技人才和数据挖掘和知识发现人才以及网络设施与商业应用经营管理人才。

参 考 文 献

[1] Big Data. Nature (http://www.nature.com/news/specials/bigdata/index.html), Sep 2008.

[2] D Agrawal, P Bernstein, E Bertino, et al. Challenges and Opportunities with Big Data — A community white paper developed by leading researchers across the United States. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf, Mar 2012.

[3] 孟小峰,慈祥. 大数据管理：概念、技术与挑战. 计算机研究与发展,2013.

[4] 赵国栋,易欢欢,糜万军,等. 大数据时代的历史机遇：产业变革与数据科学. 北京：清华大学出版社,2013.

[5] 维克多. 迈尔-舍恩伯格,肯尼思. 库克. 大数据时代：生活、工作与思维的大变革. 盛杨燕,周涛,译. 杭州：浙江人民出版社,2013.

[6] J Manyika, M Chui, B Brown, et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.

[7] 维克托・迈尔-舍恩伯格. 大数据时代. 周涛,译. 杭州：浙江人民出版社,2012.

[8] Bigdata: Thenext frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, 2013.

[9] BigData, Big ImpactNew Possibilities for International Development. http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development.

[10] Apache Hadoop. http://Hadoop.apache.org/

[11] Manish Goyal, Maryanne Q Hancock, and Homayoun Hatami. Selling into Micromarkets. Harvard Business Review, 2012.

[12] [瑞士]亚历山大・奥斯特瓦德,[比利时]伊夫・皮尼厄. 商业模式新生代. 王帅,等译. 北京：机械工业出版社,2011.

[13] 王琴. 基于价值网络重构的企业商业模式创新. 中国工业经济,2011.

[14] 黄升民,刘珊."大数据"背景下营销体系的解构与重构. 现代传播,2012,(11).

[15] 赵勇,林辉,沈寓实等. 大数据革命——理论、模式与技术创新. 北京：电子工业出版社,2014.

[16] 赵勇. 架构大数据——大数据技术与算法解析. 北京：电子工业出版社,2015.

[17] DAMA 数据管理知识体系指南. 马欢,刘晨,等译. 北京：清华大学出版社,2012.

[18] 数据治理体系——中国移动企业级大数据平台系列规范. 2014.

[19] http://www.ibm.com/developerworks/cn/bigdata/governance/index.html. Phillip Russom, Managing Big Data, TDWI Best Practices Report, Fourth Quarter 2013.

[20] 孟小峰,慈祥. 大数据管理：概念、技术与挑战. 计算机研究与发展,2013,(01).

[21] 苗放. 面向数据的安全体系结构初步研究. 中心通讯技术,2016,(1),10.3969/j.issn.1009-6868.2016.01.005.

[22] [美]Rachel Schutt,[美]Cathy O'Neil. Doing Data Science. 南京：东南大学出版社,2014.

[23] 黄德才. 数据仓库与数据挖掘教程. 北京：清华大学出版社,2016.

[24] 李志刚,马刚. 数据仓库与数据挖掘的原理及应用. 北京：高等教育出版社,2008.

[25] 孔芳,钱雪忠. 关联规则挖掘中对 Apriori 算法的一种改进研究. 计算机工程与设计,2008,29(16)：4220-4221.

[26] 李清峰,杨路明,张晓峰,等. 数据挖掘中关联规则的一种高效的 Apriori 算法. 计算机应用与软件,2004,21(12)：84-86.

[27] 杨晓平.关联规则 Apriori 算法的改进.浙江海洋学院学报,2006,25(2):176-182.
[28] Pang-Ning Tan,Michael Steinbach,Vipin Kumar.数据挖掘导论.北京:人民邮电出版社,2006.
[29] Han JW,Kamber M. Data mining:concept and techniques. Morgan Kanfmann,San Mateo,2000.
[30] Girolami M,Mercer kernel-based clustering in feature space,IEEE Trans on Neural Networks,2002,13:780-784.
[31] Gary K,Honaker J,Joseph A,Scheve K. Listwise deletion is evil: What to do about missing data in political science. 2000. http://Gking. Harvard. edu
[32] 沈红斌,杨杰,王士同.基于信息理论的合作模糊聚类算法研究.计算机学报,2005,8:1287-1294.
[33] Wen Zhang, Taketoshi Yoshida, Xijin Tang, et al. Text clustering using frequent itemsets. Knowledge-Based Systems,2010,23(5):379-388.
[34] Hong-Bin Shen,Jie Yang,XiaoJun Liu,et al. Using Supervised Fuzzy Clustering to Predict Protein Structural Classes,Biochemical and Biophysical research,2005,334:577-581.
[35] Leski J. Towards a robust fuzzy clustering,Fuzzy Sets and Systems,2003,137:215-233.
[36] 阎红灿.本体建模与语义 Web 知识发现.北京:清华大学出版社,2015.
[37] A Topchy,A K Jain,W Punch. A Mixture Model of Clustering Ensembles. In: Proceedings of the SIAM International Conference on Data Mining,Lake Buena Vista,Florida,2004:22-24.
[38] Shi Zhong,Joydeep Ghosh. A Unified Framework for Model-based Clustering. Journal of Machine Learning Research,2004,4(6):1001-1038.
[39] P Ponmuthuramalingam,T Devi. Effective Term Based Text Clustering Algorithms. International Journal on Computer Science and Engineering,2010,2(5):1665.
[40] Ralph Kimball.数据仓库工具箱:维度建模的完全指南(第二版).北京:电子工业出版社,2003.
[41] Wolf-Tilo Balke. Introduction to Information Extraction: Basic Notions and Current Trends. Datenbank-Spektrum,2012,12(2):81-88.
[42] 付年钧,彭昌水,王慰.中文分词技术及其实现.软件导刊,2011.
[43] 张启宇,朱玲,张雅萍.中文分词算法研究综述.情报探索,2008,11:53-56.
[44] Cui-xia Li,Nan Lin. A Novel Text Clustering Algorithm. Energy Procedia,2011,13:3583-3588.
[45] Xinwu Li. A New Text Clustering Algorithm Based on Improved K_means. Journal of Software,2012,7(1):95-101.
[46] Hong-Bin Shen,Jie Yang,Shi-tong Wang,Outlier Detecting in Fuzzy Switching Regression Models. AIMSA 2004:LNAI 3192. 208-215.
[47] Vapnik V N. Statistical learning Theory. Wiley,New York,1998.
[48] Nam Hun Park,Won Suk Lee. Statistical grid-based clustering over data streams. ACM SIGMOD Record,2004,33(1):32-37.
[49] Shi Zhong,Joydeep Ghosh. A Unified Framework for Model-based Clustering. Journal of Machine Learning Research,2004,4(6):1001-1038.
[50] [美]Sunll Soares.大数据治理.匡斌,译.北京:清华大学出版社,2014.
[51] [美]Jay Liebowitz.大数据与商业分析.刘斌,曲文波,林建忠,等译.北京:清华大学出版社,2015.
[52] Kui Fang,Weiqiong Bu,Wu Luo,et al. Chinese Word Segmentation for Agriculture. Journal of Software,2013,8(5):1219-1226.
[53] Xu Sun,Yaozhong Zhang,Takuya Matsuzaki,et al. Probabilistic Chinese word segmentation with non-local information and stochastic training. Information Processing and Management,2013,49(3):626-636.
[54] Chu-Ren Huang,Nianwen Xue. Words without Boundaries: Computational Approaches to Chinese Word Segmentation. Language and Linguistics Compass,2012,6(8):494-505.
[55] Hamish Cunningham. Developing Language Processing Components with GATE Version 7(a User

Guide) For GATE version 7.1-snapshot(development builds)(built March 22,2012).
[56] IBM 智慧医疗解决方案
[57] 全国环境统计公报(2014 年)http://www.zhb.gov.cn/gzfw_13107/hjtj/qghjtjgb/201605/t20160525_346106.shtml
[58] 宋伟民.中国雾霾天气的现状与挑战.环境与健康展望.2013,8(12):3-4.
[59] 廖伟明.中国的大气污染及治理概况.http://www.docin.com/p-3 17759803.html
[60] 吴晓青.我国大气氮氧化物污染控制对策.环境保护,2009,(16):9-11.
[61] 李超慈.基于电力二氧化硫排放总量控制的燃煤电源规划模型研究.北京:华北电力大学,2010.
[62] 郑艳琳,李福利.燃煤二氧化硫排放量的回归测算模型.能源环境保护,2009,23(3):47-50.
[63] Cofala J,Amann M,Gyarfas F,et al. Cost-EffectiVe Contml of S02Emissions in Asia. Joumal of Environmental Management,2004,72(3):149-161.
[64] 王青,邹骥,王磊.区域二氧化硫控制优化模型研究.中国人口资源与环境,2005,15(6):67-70.
[65] Gao C,Yin H,Ai N,et al. Historical analysis of S02 pollution control policies in China. Environmental management,2009,43(3):447-457.
[66] 秱小洛,曹国良,黄学敏.中国区域氮氧化物排放清单.环境与可持续发展,2008.
[67] 王丽琼.中国氮氧化物排放区域差异及减排潜力分析.地理与地理信息科学,2010.
[68] 房晟忠,赵世民,李发荣,等.氮氧化物排放模型及排放清单研究现状.环境科学导刊,2010.
[69] 郭斌,廖宏楷,徐程宏,等.我国 SCR 脱硝成本分析及脱硝电价政策探讨.热能动力工程,2010,25(4):437-440.
[70] 马忠丽,王科俊,莫宏伟,等.基于免疫遗传算法的环境经济负荷调度.电力系统及其自动化学报,2006.
[71] 余欣梅,熊信艮,吴耀武,等.电力系统调峰电源规划优化模型探讨及其应用.中国电力,2003.
[72] 张文泉,董福贵,张世英,等.电厂鲁棒性组合研究.华东电力,2003.
[73] 蔡明坤.装有脱硝系统锅炉用回转式预热器设计存在问题和对策.锅炉技术,2005.
[74] 周建国,崔冰,赵毅.基于外部性内部化与国家意愿支付的脱硝电价定价研究.技术经济,2010.
[75] 黄少忠,左源,赵立志.关于脱硝电价政策的研究和建议.国家电力监管委员会,2012.
[76] 周春静.基于经营期电价的火电厂脱硝电价定价研究与措施建议.北京:华北电力大学,2011.
[77] 周建国,安园园,段三良,等.基于外部性与动态盈亏平衡的燃煤电厂脱硝电价研究.电力系统保护与控制,2010.
[78] 王若颖.XBP 电厂 300MW 机组脱硝改造工程方案设计与效益评价.北京:华北电力大学,2013.
[79] 王璐.华能九台电厂脱硝改造项目可行性研究.长春:吉林大学,2013 .
[80] 张新立.燃煤电厂老机组脱硝改造技术探讨与分析.北京:华北电力大学,2013.
[81] 郑永亮.NJCRTP 燃煤发电机组脱硝改造项目可行性研究.南京:南京理工大学,2012.
[82] 中国金融四十人论坛课题评审会暨第 64 期"双周圆桌"内部讨论会纪要.互联网金融模式与未来金融业发展.新金融评论,2012,(1).
[83] 谢平,邹传伟.互联网金融模式研究.金融研究,2012,(12).
[84] 方方."大数据"趋势下商业银行应对策略研究.新金融,2012,(12).
[85] 巴曙松,湛鹏.互动与融合:互联网金融时代的竞争新格局.2012,(12).
[86] 金融互联网化相互包容共生——中国银行业协会互联网金融研讨会纪要.NEW FINANCE,2013,(12).
[87] 吴澄,孙优贤,王天然,等.信息化与工业化融合战略研究.北京:科学出版社,2013.
[88] 范玉顺.i 时代信息化战略管理方法.北京:清华大学出版社,2015.
[89] 郭重庆.互联网将重新定义制造业.东沙湖论坛中国管理百人报告会,2014.
[90] 国务院发展研究中心.中国制造 2025.2014.
[91] 德国人工智能研究中心.德国工业 4.0 研究报告.2014.

[92] 周宏仁.大力推进信息化与工业化的融合.中国电子报,2010.
[93] 顾新建,纪杨建,祁国宁.制造业信息导论.杭州:浙江大学出版社,2010.
[94] 李晨晖,崔建明,陈超泉.大数据知识服务平台构建关键技术研究.情报资料工作,2013.
[95] 宗威,吴锋.大数据时代下数据质量的挑战.西安交通大学学报(社会科学版),2013.
[96] 胡雄伟,张宝林,李抵飞.大数据研究与应用综述(下).标准科学,2013.